ANONOS

Schrems II Compliant Supplementary Measures

11 November 2020

ANONOS

Summary

The attached documents serve as evidence of the feasibility and practicability of the proposed measures enumerated in the EDPB's *Recommendations 01/2020 on Measures That Supplement Transfer Tools to Ensure Compliance With the EU Level of Protection of Personal Data;* the first document, prepared by Anonos, describes several use cases, and the others include independent audits, reviews and certifications of Anonos state-of-the-art technology that leverages European Union Agency for Cybersecurity (ENISA) recommendations for GDPR compliant Pseudonymisation to enable EDPB proposed measures.

Table of Contents

1. Maximizing Data Liquidity - Reconciling Data Utility & Protection **3**
2. IDC Report - Embedding Privacy and Trust Into Data Analytics Through Pseudonymisation **20**
3. Data Scientist Expert Opinion on Variant Twins and Machine Learning **29**
4. Data Scientist Expert Opinion on BigPrivacy **38**

MAXIMIZING DATA LIQUIDITY

RECONCILING DATA UTILITY & PROTECTION

August 2020

www.anonos.com

EXECUTIVE SUMMARY

Eight years of research and development have created a solution that optimizes both data protection and data use to **maximize data liquidity** lawfully & ethically.

Anonos BigPrivacy technology transforms data into state-of-the-art secure Variant Twin data assets. Variant Twins embed controls into the data to eliminate conflicts between data use and data protection, enabling the highest levels of accuracy and utility. This is in stark contrast to outdated approaches that subscribe to the adage that "you can have data utility or data protection, but you can't have both." Anonos resolves this issue without any compromise.

Anonos patented technology maximizes global opportunities for sharing, combining and enhancing data, to achieve the highest lawful and ethical data value while ensuring uninterrupted data access and use across any ecosystem.

- Prior to the pandemic, evolving data protection laws highlighted the limitations of compliance, security and privacy solutions to overcome the inability of consent and encryption to make desired data use lawful.[1]
- With the pandemic, in-person collection and use of data has now become impossible. Companies must rely on cloud and SaaS services for timely insights about partner and customer ecosystems.
- Effective 16 July 2020, the Schrems II invalidation of the Privacy Shield makes many cloud and SaaS services involving international data transfer unlawful.

It is a critical requirement for companies and their leaders to seek technologies to overcome these limitations so that desired data uses in the cloud and via SaaS are lawful and ethical.

[1] See World Economic Forum whitepaper, *Redesigning Data Privacy: Reimagining Notice & Consent for Human-Technology Interaction,* at https://www.weforum.org/reports/redesigning-data-privacy-reimagining-notice-consent-for-humantechnology-interaction. See also https://techcrunch.com/2020/08/14/oracle-and-salesforce-hit-with-gdpr-class-action-lawsuits-over-cookie-tracking-consent/

Use Cases: Maximizing Data Utility and Data Protection

The seven use cases below represent different business requirements for maximizing the utility of data for secondary processing. As more fully described below, Use Case 1 (Clear Text Processing) and Use Case 2 ("Anonymisation") have higher levels of data utility but at the cost of a significant risk of violating data protection requirements such as the GDPR and CCPA. Conversely, Use Case 3 (Summary Statistics) has a reduced risk of breaching data protection requirements but at the cost of reduced data utility.

Use Cases 1 – 3 are generally achievable using alternative solutions and approaches; however, Anonos patented technology uniquely enhances utility in Use Case 2. **For the other four Use Cases (4 – 7), only Anonos patented technology delivers both maximum data utility and maximum data protection enabling more timely insight and business value from data sharing, combining and enriching for Analytics: Big Data, AI, ML and BI.** No other vendor can deliver this outcome.

USE CASES		DELIVERS DESIRED UTILITY AND NECESSARY COMPLIANCE*	
		ANONOS	ALL OTHER SOLUTIONS
1	Clear Text Processing	✓	✓
2	"Anonymisation"	✓	✓
3	Summary Statistics	✓	✓
4	Big-Data, AI, ML and BI	✓	✗
5	General Purpose Analytics / Schrems II Additional Safeguards	✓	✗
6	Microsegment-based Internal and External Data Sharing	✓	✗
7	Unique, Random, Row-Level Pseudonyms for Internal Data Sharing	✓	✗

AT-RISK | MITIGATED RISK | NO EXPECTED RISK

*In compliance with footnote 2 and endnotes ii - xxvi

Each of these seven Use Cases (which will be explored in more detail, and correspond to the numbered locations 1 – 7 on the Anonos Privacy Engineering Framework, below) is labelled with one of the following levels of risk, indicating the **Anonos Compliance Guarantee**[2] with respect to the GDPR and CCPA:

- **AT RISK**: May not satisfy GDPR or CCPA requirements without additional measures.
- **MITIGATED RISK**: Reasonable evidence exists to support the use case meeting GDPR or CCPA requirements, as applicable. However, relevant regulators could rule against it.
- **NO EXPECTED RISK**: Leverages express statutory benefits under GDPR or CCPA.

Anonos Privacy Engineering Framework

The following graphic and subsequent discussion highlights the Privacy Engineering considerations associated with these use cases, and how Anonos' unique state-of-the-art Privacy Engineering Tools can be configured to maximize data liquidity by simultaneously optimizing data protection and data utility. This process is also in compliance with today's regulatory landscape against the full range of potential use cases, and the privacy protection landscape of tomorrow.

---

[2] **Anonos Compliance Guarantee**: The GDPR is expressly technology neutral (e.g., Recital 15 states "In order to prevent creating a serious risk of circumvention, the protection of natural persons should be technologically neutral and should not depend on the techniques used") and the CCPA does not require specific technology. Endnotes ii – xvii highlight express statutory benefits under the GDPR and endnotes xviii – xxvi highlight express statutory benefits under the CCPA for, inter alia, enforcing Functional Separation using GDPR-compliant Pseudonymisation or CCPA-compliant heightened De-Identification. As a technology vendor, Anonos cannot guarantee the proper use by data controllers, processors or third parties of its technology. However, Anonos will guarantee that if Anonos technology is used correctly in compliance with provided documentation and all applicable laws, rules and regulations: the technology will satisfying the statutory provisions noted in: (a) endnotes ii – xvii with regard to the GDPR, plus Article 25 Data Protection by Design and by Default requirements; and (b) endnotes xviii-xxvi with regard to the CCPA. Anonos also assists data controllers, processors and third parties using Anonos technology with applicable reporting, audit and disclosure obligations related to technical and organizational measures including GDPR Articles 28(3)(h), 30(1)(g), 30(2)(d), 32(1), and 35(7)(d).

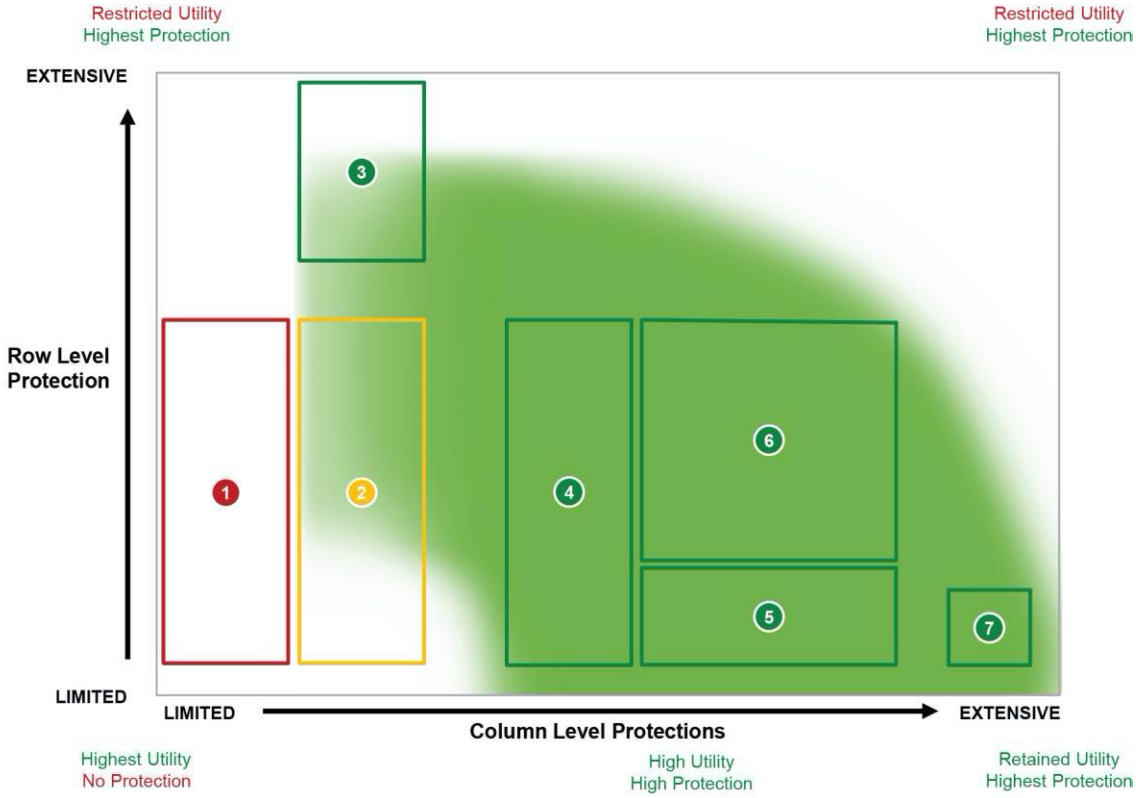
ANONOS PRIVACY ENGINEERING FRAMEWORK

APPLICATION AGAINST SEVEN USE CASES

100% of the Protection & Utility Design Space

Maximized High Protection / High Utility Target Zone

A OUTPUT RECORDS	B K ANONYMITY
1 record	k = N
N > x > 1 records	k >> 5
	k > 5
	k = 5
	1 < k < 5
All N records	k = 1



C INCLUSION IN OUTPUT	All Fields		Some Fields				1 Field
D COLUMNS TREATED	None	Direct Identifiers	Direct Identifiers Indirect Identifiers		Direct Identifiers Indirect Identifiers Attributes		All
E PRIVACY ACTIONS	Clear Text	Masking Generalization Aggregation	Associative Directly Reversible Pseudonyms	Associative Controllably Relinkable Pseudonyms	Random Directly Reversible Pseudonyms	Random Controllably Relinkable Pseudonyms	Row Level Pseudonyms

BIGPRIVACY USE CASE DESIGN SPACES

1. Clear Text Processing
2. "Anonymisation"
3. Summary Statistics
4. Big-Data, AI, ML and BI
5. General Purpose Analytics / Schrems II Additional Safeguards
6. Microsegment-based Internal and External Data Sharing
7. Unique, Random, Row-Level Pseudonyms for Internal Data Sharing

Anonos BigPrivacy patented technology integrates the application of a full range of established and novel accuracy-preserving privacy engineering techniques across five dimensions (represented by A – E, above) simultaneously, and independently from one another:

A. Restriction of output to a subset of input records
B. Restriction of output based on minimum record counts in subgroups
C. Restriction of output to specific columns
D. Application of protection to direct identifiers, indirect identifiers, and attributes
E. Application of anonymization techniques such as masking, generalization, and aggregation; and application of four different types of field-level pseudonyms

The range and flexibility enabled by BigPrivacy provides privacy engineers access to a larger design space, allowing the creation of fit-for-purpose, privacy-respectful data assets (called Variant Twins) that meet the requirements for utility and data protection in use not previously obtainable. This enables compliant use of data for a more extensive range of high-value use cases. For use cases in which minimal loss of accuracy relative to clear text is a requirement, significantly far higher levels of utility and protection can now be achieved.

For example, machine learning requires granular record level data while keeping as many fields, and using as little aggregation or generalization, as possible. BigPrivacy enables the creation of Variant Twins that mask or omit direct identifiers and aggressively pseudonymize categorical variables while leaving numerical attributes in clear text or only modestly generalized. When used in machine learning models including classification trees, neural nets, and regression, Variant Twins deliver results with 99%+ accuracy and utility, when compared to clear text processing.[3] Note that all the above five dimensions can be traversed independently. For example, on the vertical axis, the output might be restricted to include only a few of the records (via a filter applied to one field, for example “age” > 70 years) but k-anonymity could be set at k=1. Similarly, on the horizontal axis, only a handful of fields might be included in the output, and direct and indirect identifiers could be treated using generalization and random controllably linkable pseudonyms, with row-level pseudonyms included in the output.

[3] See https://www.anonos.com/data-scientist-expert-opinion

Use Case 1: Clear Text Processing

Maximum data utility but no incremental protection of data in use relative to the source data itself. May be appropriate for internal use cases under controlled conditions.

- **Specification:** All rows are used to create output, k is commonly set to 1, but could be as high as 50 or even 100, all fields are in output, no fields are treated, all fields are clear text.
- **Output**: This results in an output that is a complete (if k=1) or nearly complete duplicate of the source data.
- **Risk Level**: **AT RISK**

Use Case 2: "Anonymisation"

Often used for General-Purpose Analytics as well as Big Data, AI, ML and BI applications. Though frequently referred to as "anonymous" data, the output would not be "anonymous" as defined under the GDPR nor "de-identified" as defined under the CCPA. This is because reidentification is often achievable when the data is combined with other data sets. If this is the case, then output would not be exempt from GDPR or CCPA jurisdiction. Utility is high, but protection in use is limited.

- **Specification**: All or some records are used to create output via filtering, k could be as low as 1, more commonly at 5, and could be as high as 50 or even 100, direct identifiers are removed, no treatment of indirect identifiers, numerical attributes are lightly generalized.
- **Output**: Output may be referred to as "anonymous" or "deidentified," however, its use should be limited to localized applications under controlled conditions. This is because the output will be susceptible to unauthorized re-identification when combined with external data.
- **Anonos Value-Add:** Unlike traditional anonymisation or deidentification techniques, BigPrivacy can create a comparable pseudonymised output which is not subject to unauthorized re-identification, but which can be relinked to the original source data by authorized parties.
- **Risk Level:** **MITIGATED RISK**

Use Case 3: Summary Statistics

Some protection of data in use, particularly for numerical attributes, such as in a classic "summary statistics" report. Some data utility preserved, but nuances and accuracy are reduced due to the loss of record-level granularity resulting from aggregation

- **Specification:** All records are used to create output, k is effectively much greater than 5 due to aggregation in calculating summary statistics, some fields are included in the output (typically some indirect identifiers and most numerical attributes).
- **Output:** All output is clear text, but all numerical attributes are aggregated to statistics such as averages, min/max, standard deviations, etc.
- **Risk Level:** **NO EXPECTED RISK**

Use Case 4: Big Data, AI, ML and BI

This use case delivers extremely high resistance to re-identification while still preserving high utility for Big Data, AI, ML and BI applications, both within and between organizations. It will return results that are essentially identical to using clear text. In addition, directly reversible pseudonyms facilitate model interpretability.

- **Specification:** All or some records are used to create output via filtering, k is frequently set to 1, but could be as high as 50 or even 100 in very large data sets,, direct identifiers are removed, extensive treatment of indirect identifiers, all categorical indirect identifiers and attributes are pseudonymised using associative directly reversable pseudonyms, attributes are generalized where possible without reducing utility.
- **Output:** This output results in a GDPR-compliant pseudonymised (CCPA-compliant deidentified) data set comprising pseudonymised indirect identifiers and categorical attributes, and clear text numerical attributes.
- **Anonos Value-Add:**
    - Allows Big Data, AI, ML and BI data sets that comply with GDPR pseudonymisation requirements/CCPA de-identification requirements. This enables lawful and ethical data sharing, enrichment and combining with results that are essentially identical to using clear text, while providing dramatically improved protection of data in use. In addition, directly reversible pseudonyms facilitate model interpretability.

  - Improves predictability of operations under GDPR and CCPA for Withdrawal of Consent/Right to be Forgotten and Withdrawal of Consent/Deletion Requests, respectively.
- **Risk Level:** NO EXPECTED RISK (Note: due to the use of algorithmically derived "associative directly reversable pseudonyms" versus randomly assigned "associative controllably relinkable pseudonyms" as in Use Case 5 below, Use Case 4 may be held not to satisfy Schrems II requirements for supplemental measures for international EU personal data transfer. For this reason, it may be advisable to use randomly assigned "associative controllably relinkable pseudonyms" as described in Use Case 5 below to ensure compliance with Schrems II requirements for international EU personal data transfer).

Use Case 5: General Purpose Analytics / Schrems II Additional Safeguards

This is suitable for general purpose analytics and many types of data sharing and enrichment within and between organizations.

- **Specification**: All or some records are used to create output via filtering, could be as low as 1, though more typically set to 5 and could be as high as 50 or even 100, direct identifiers are removed, extensive treatment of indirect identifiers, all categorical indirect identifiers and attributes are pseudonymised using associative controllably relinkable pseudonyms, numerical attributes lightly generalized as necessary while still preserving analytical utility.
- **Output**: This output is a more aggressively protected GDPR-compliant pseudonymised (CCPA-compliant de-identified) data set due to pseudonyms that are controllably relinkable rather than directly reversible.
- **Anonos Value-Add:**
  - Unlike traditional anonymisation and deidentification techniques, BigPrivacy can create comparable pseudonymised output which is not subject to unauthorized reidentification, but which can be relinked to the original source data by authorized parties.
  - With appropriate organizational controls regarding access to the information necessary for relinking, this approach can be used to meet Schrems II requirements. Schrems II requires additional safeguards to supplement contractual provisions (Standard Contractual Clauses (SCCs) between parties or Binding Corporate Resolutions (BCRs) within a global organization) for transfers of EU

data to cloud and SaaS providers headquartered in non-EU countries that have not received an EU Commission adequacy determination.

- Improves predictability of operations under GDPR and CCPA for Withdrawal of Consent/Right to be Forgotten and Withdrawal of Consent/Deletion Requests, respectively.

- **Risk Level:** NO EXPECTED RISK

Use Case 6: Microsegment-based Internal and External Data Sharing

This approach is be suitable for data sharing, enrichment and combining within and between organizations where regulations or policies prevent using identity resolution approaches.

- **Specification:** All or some records are used to create output via filtering, k=5 or higher, direct identifiers are removed, extensive treatment of indirect identifiers, all categorical indirect identifiers and attributes are pseudonymised using associative controllably relinkable pseudonyms, numerical attributes are generalized to summary statistics using microsegments defined by combinations of indirect identifiers.
- **Output:** This is a very aggressively protected data set making use of both controllably linkable pseudonyms and significant aggregation that is extremely privacy respectful.
- **Anonos Value-Add:**
  - Enables data sharing between unrelated organizations under even the most restrictive data privacy regulations (GDPR, CCPA) and data sovereignty/localization laws.
  - With appropriate organizational controls regarding access to the information necessary relinking, this approach can be used to meet Schrems II requirements. Schrems II requires additional safeguards to supplement contractual provisions (Standard Contractual Clauses (SCCs) between parties or Binding Corporate Resolutions (BCRs) within a global organization) for transfers of EU data to cloud and SaaS providers headquartered in non-EU countries that have not received an EU Commission adequacy determination.
  - Improves predictability of operations under GDPR and CCPA for Withdrawal of Consent/Right to be Forgotten and Withdrawal of Consent/Deletion Requests, respectively.
- **Risk Level:** NO EXPECTED RISK

Use Case 7: Unique, Random Row-Level Pseudonyms Internal Data Sharing

This approach would be used for the exchange of information between two parties inside the same organization in a secure fashion. The approach avoids passing any source data itself, since the receiving party can access the source data later if needed and authorized.

- **Specification**: All records are used to create output, k=1, no fields are included in the output other than row-level pseudonyms.
- **Output**: This output will contain a single, unique, random row-level pseudonym for each source record.
- **Anonos Value-Add:**
    - Because these pseudonyms contain no information derived from the source data, the output ensures the maximum possible protection.
    - Additionally, this approach avoids passing any source data itself, since the receiving party can access the source data later if needed and authorized.
    - Improves predictability of operations under GDPR and CCPA for Withdrawal of Consent/Right to be Forgotten and Withdrawal of Consent/Deletion Requests, respectively.
- **Risk Level:** NO EXPECTED RISK

ANONOS ADDRESSES CURRENT & FUTURE BUSINESS OBJECTIVES

Business Use Cases	Simultaneous Delivery of Required Utility and Protection [i]
Analytics: Big Data, AI, ML & BI	
Internal Teams & Internal Algorithms: • Department level single assets • Across department multiple assets • Across business unit's multiple departments • Employee data access • Across company / international transfer (e.g., offshore central analytics teams)	**Yes**
External Partners & 3rd Party Algorithms: • Department level single assets • Across department multiple assets • Across business unit's multiple departments • Across company / international transfer (e.g., offshore central analytics teams)	**Yes**
Combining & Enrichment: Marketing & Client Engagement	
Internal & Client Data Sets: • Combining client data from different departments • Single view of the client • Next step activity to drive client engagement (CRM / Web) • Enriching client data from multiple internal sources • Deliver the value of the data lake strategy	**Yes**
External Partners & 3rd Party Data: • Engaging with corporate clients around their data • Access to 3rd party data sets • Access to alternative data sources • Ability to combine & enrich with across all data sources	**Yes**
Data Monetization: New Revenue Streams	
Data as an Asset: • Create new assets with enhanced security and privacy while retaining full value for 3rd party sharing • Maximum utility in any application or ecosystems • Maximum accuracy in any application or ecosystems • Multiple and reusable data assets • Ongoing data asset enrichment	**Yes**

ANONOS ADDRESSES LEGAL & REGULATORY REQUIREMENTS

Key Data Protection Principles Delivered by Anonos	Simultaneous Delivery of Required Utility and Protection
General Data Protection Regulation (GDPR)	
1. **Functional Separation**[ii]	**Yes**
2. **GDPR-Compliant Pseudonymisation**[iii]	**Yes**
3. **Legitimate Interest Processing**[iv]	**Yes**
4. **Lawful International Data Transfer and Cloud-Based (SaaS) Processing after Schrems II Invalidation of Privacy Shield and Need for Supplemental Measures**[v]	**Yes**
5. **Predictability of Operations**[vi]	**Yes**
Benefits of GDPR-Compliant Pseudonymisation	
6. Legitimate Interest Processing[vii]	**Yes**
7. Secondary Processing (change of purpose / further processing)[viii]	**Yes**
8. Data Minimisation[ix]	**Yes**
9. Storage[x]	**Yes**
10. Security[xi]	**Yes**
11. Profiling[xii]	**Yes**
12. Sharing, Combining and Enhancing[xiii]	**Yes**
Benefits of Legitimate Interest Processing	
13. Limit the right to restrict processing[xiv]	**Yes**
14. Limit the right to data portability[xv]	**Yes**
15. Limit the right to object[xvi]	**Yes**
16. Resolve clinical trial consent conflicts[xvii]	**Yes**
California Consumer Privacy Act (CCPA) and Other Evolving Data Protection Laws	
17. **Functional Separation**[xviii]	**Yes**
18. **CCPA Heightened De-Identification**[xix]	**Yes**
19. **Predictability of Operations**[xx]	**Yes**
Benefits of CCPA Heightened De-Identification	
20. Research[xxi]	**Yes**
21. Collection[xxii]	**Yes**
22. Use[xxiii]	**Yes**
23. Retention[xxiv]	**Yes**
24. Sale[xxv]	**Yes**
25. Disclosure[xxvi]	**Yes**

[i] See https://www.anonos.com/data-scientist-expert-opinion

[ii] The Article 29 Working Party ("WP29"), and the European Data Protection Supervisor ("EDPS") recommend the use of "Functional Separation." Functional Separation involves separating information value from identity to enable the discovery of trends and correlations independent from any subsequent authorised application of the insights gained to the individuals concerned. In *Opinion 03/2013 on Purpose Limitation,* the WP29 highlighted the "prominent role in our analysis for different kinds of safeguards, including technical and organizational measures to ensure functional separation, such as full or partial anonymisation, pseudonymisation, aggregation of data, and privacy-enhancing technologies." (See https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf at pages 13, 26, 27, 29, 30, 31, 32, 33, 40, and 46.) In addition, a 2015 report by the EDPS, *Meeting The Challenges of Big Data – A Call For Transparency, User Control, Data Protection By Design And Accountability*, highlighted functional separation as a potential solution to help resolve conflicts between innovative data use and data protection. (See https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf at page 15. Additional information on functional separation is available at www.MosaicEffect.com). Under the GDPR, the concept of Functional Separation is embodied in the Article 4(5) definition of "Pseudonymisation" which requires that the information value of data must be separated from the identity of data subjects and that additional securely stored information must be necessary to reidentify data subjects, and then only under controlled conditions.

[iii] **It is critical to note that under the GDPR, Pseudonymisation is now defined as an outcome and not a technique.** Our experience is that the existence, and more importantly, the significance of this change, is not well appreciated outside the EU regulatory community. Over time we anticipate that the foresight in making this fundamental change will become evident due to its broad utility and effectiveness in resolving conflicts between innovative data use and data protection. Before the GDPR, Pseudonymisation was widely understood to mean replacing direct identifiers with tokens and was applied to individual fields independently within a data set. It was merely a Privacy Enhancing Technique ("PET"). With the elevation of Pseudonymisation now to an outcome, to achieve GDPR-compliant Pseudonymisation, it has become necessary to protect not only direct identifiers but also indirect identifiers. In addition, instead of being applied only to individual fields, GDPR-defined Pseudonymisation, in combination with the GDPR definition for Personal Data, now requires that the outcome must apply to a data set as a whole (the entire collection of direct identifiers, indirect identifiers and other attributes), and consideration must be given to the degree of protection applied to all attributes in a data set. Finally, the preceding must be accomplished while still preserving the data's utility for its intended use. As a result, pre-GDPR approaches (using a static token on a direct identifier, which unfortunately is still widely and incorrectly referred to as "pseudonymisation") will rarely, if ever, meet the heightened GDPR requirements of Pseudonymisation.

[iv] Under the GDPR, consent is available as a legal basis only if the desired processing can be described in advance with sufficient specificity so that data subjects can provide knowing and voluntary consent at the time of initial data collection (see GDPR Articles 4(11) and 6(1)(a)). It is not possible to secure GDPR-compliant consent for processing that occurs in the future, which cannot be described adequately at the time of data collection. Reconsenting is necessary each time the processing changes from the specific use disclosed at the time of collection. IDC reports that a top European financial services firm obtained at best a 60% success rate each time it tried to obtain reconsent from customers. "Bundling" of consent to get approval for analytics, AI or ML within the acceptance of general terms and conditions, or "tying" the provision of a contract or a service to receiving consent to process analytics, AI or ML, is not lawful because consent in these circumstances is not freely given (see GDPR Recital 43 and Article 7(4) and https://techcrunch.com/2018/10/03/europe-is-drawing-fresh-battle-lines-around-the-ethics-of-big-data/.) Data subjects can only lawfully consent to data uses that are explicitly explained at the time of providing consent (see GDPR Recital 32). **Organizations can overcome the limitations of consent for lawful analytics, AI and ML by using GDPR-compliant Pseudonymisation to support Legitimate Interest processing(see GDPR Article 6(1)(f)) to (a) enable processing that cannot be described with required specificity at the time of initial data collection; and (b) avoid having to seek reconsent each time different processing of data is desired**. GDPR-compliant Legitimate Interest processing requires more than mere claims of having a "legitimate interest" in the outcome of processing. To serve as a valid legal basis, Legitimate Interest processing must satisfy a three-part test; the first two tests are relatively easy to meet while the third test requires technical and organizational safeguards. The three tests are: (a) Legitimate Interest test - is there a legitimate interest behind the processing; (b) Necessity test - is the desired processing necessary for that purpose; and (c) Balancing of Interest test - do technical and organizational safeguards counterbalance the interests of the data controller (or a third party) against data subjects' rights and freedoms. **Technical and organizational safeguards that can "play a role in tipping the balance in favour of the controller" include: (a) Functional Separation and (b) Pseudonymisation** (see WP217 at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf, page 42).

[v] See https://www.anonos.com/variant-twin-value-proposition

[vi] Consent exposes data controllers to unpredictable interruptions in processing - e.g., when consent is revoked by a data subject or they exercise their Article 17 Right to be Forgotten/Right to Erasure. Article 11 provides an exemption from these rights if "the controller is able to demonstrate that it is not in a position to identify the data subject." Since the GDPR does not require a controller to hold additional information "for the sole purpose of complying with this Regulation," a data controller can use Anonos' patented pseudonymisation-enabled Controlled Linkable Data and delete information that would identify individual data subjects. The concept of Controlled Linkable Data was presented at an International Association of Privacy Professionals (IAPP) program titled *How to Comply with the GDPR While Unlocking the Value of Big Data* featuring Gwendal Le Grand, Director of Technology and Innovation at the French Data Protection Authority - the CNIL, Mike Hintze, Partner at Hintze Law and former Chief Privacy Counsel and Assistant General Counsel at Microsoft, and Gary LaFever, CEO and

General Counsel at Anonos and former law partner at Hogan Lovells (see https://www.anonos.com/iapp-gdpr-data-analytics-webinar-replay) and explained in a White Paper co-authored by Messrs. Hintze and LaFever titled *Meeting Upcoming GDPR Requirements While Maximizing the Full Value of Data Analytics.* See https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2927540

[vii] See discussion regarding Legal Principle #3 above. See also Articles 5(1)(a), 6(1)(f), and WP217 at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

[viii] Under Article 5(1)(b), when an organization processes personal data obtained for a particular permitted purpose, it cannot process it further except for compatible purposes. However, "further processing" of personal data may be deemed compatible with the original purpose if the processing satisfies the requirements of: (a) Article 6(4) with respect to further "processing for a purpose other than that for which the personal data have been collected…not based on the data subject's consent" – for example, where processing is based on Legitimate Interest processing; or (b) Article 89(1) with respect to processing conducted for "archiving purposes in the public interest," "scientific or historical research purposes," or "statistical purposes." The GDPR highlights Pseudonymisation as a safeguard to help ensure that such further processing is lawful. For example: (a) Article 89(1) specifically highlights pseudonymisation and notes that further processing for "archiving purposes in the public interest," "scientific or historical research purposes," or "statistical purposes" is specifically considered not to be incompatible with the initial purposes if appropriate safeguards for data subjects are provided to ensure, in particular, data minimisation; and (b) pseudonymisation is also explicitly recognised as a safeguard under Article 6(4)(e) to help ensure that further processing of personal data "[are] compatible with the purpose for which the personal data are initially collected" in compliance with Article 5(1)(b) ("purpose limitation") requirements.
See also WP203 at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

[ix] See Articles 5(1)(c) and 89(1)

[x] See Articles 5(1)(e) and 89(1)

[xi] See Articles 5(1)(f) and 32

[xii] See Recital 71, Article 22 and WP251 at https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=49826

[xiii] The GDPR clarifies and enhances the privacy rights of individual data subjects with new well-known rights such as the "right to be forgotten," the "right to data portability" and more. However, under GDPR Articles 11(2) and 12(2), if the purposes for which an organization processes personal data do not or no longer require identification of an individual, and the organization can show that it is not in a position to identify the data subject, then it does not need to comply with these data subject rights (subject to the right of a data subject under Article 11(2), for the purpose of exercising his or her rights, to provide additional information enabling his or her identification). If personal data is pseudonymised using BigPrivacy, so that a given controller or processor cannot identify the individuals concerned, such organization may not be subject to certain obligations. BigPrivacy helps to limit the risk of using personal data by data controllers and processors down the data chain. This enables the data to be used going forward in a "risk-reduced" manner, which dramatically limits the likelihood of data subjects being reidentified. In addition, non-identifying Variant Twin versions of data processed under the lawful basis of Legitimate Interests become the proprietary data assets of an organization, with respect to which there is no obligation to provide copies to third parties (which may be competitors – e.g., FinTechs under the second Payment Services Directive (PSD2), a European directive designed to boost competition and the variety of financial services offerings). In addition, the right of data portability under GDPR Article 20 applies only to personal data processed using consent (under Article 6(1)(a) or Article 9(2)(a)) or contract (under Article 6(1)(b)), and not to data processed based on Legitimate Interests. As long as a controller can prove "compelling legitimate grounds for processing which override the interests, rights and freedoms" of data subjects due to the use of state-of-the-art technical and organizational safeguards (or "for the establishment, exercise or defense of legal claims"), objections by data subjects under Article 21 to using Variant Twin data for Legitimate Interests processing for sophisticated data analysis, AI, ML, sharing, combining, or enriching may be unsuccessful. In addition, GDPR-compliant Pseudonymisation can alleviate a data controller's requirements to carry out data subjects' rights of access under Article 15, rectification under Article 16, and erasure ("right to be forgotten") under Article 17. Article 11 provides an exemption from these rights if "the controller is able to demonstrate that it is not in a position to identify the data subject." Since the GDPR does not require a controller to hold additional information "for the sole purpose of complying with this Regulation," a data controller may use pseudonymisation techniques and subsequently delete information that would enable the reversal of the pseudonymisation to identify individual data subjects. See also WP259 at https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf

[xiv] If a data controller uses GDPR-compliant Legitimate Interests processing, under GDPR Article 18(1)(d), they do not have an obligation to comply with requests by data subjects to restrict processing if the controller's interests can be shown to prevail over the concerns of data subjects because of the existence of technical and organizational safeguards.

[xv] Under GDPR Article 20(1), data controllers using Legitimate Interests processing do not have an obligation to comply with requests by data subjects to provide data to competitors, which right only applies to consent or contract-based processing.

[xvi] Data subjects do not have the right to object to processing under GDPR Article 21(1) if a data controller uses GDPR-compliant Legitimate Interests processing and the controller's interests can be shown to prevail over the concerns of data subjects because of the existence of technical and organizational safeguards. However, data subjects always have the right under Article 21(3) to not receive direct marketing outreach resulting from data processing.

[xvii] EU clinical trial regulations deal with the withdrawal of consent by study participants with less far-reaching impacts than the separate issue of withdrawal of consent for purposes of the GDPR. The latter can potentially lead to the termination of entire studies based on a request for withdrawal by a single study participant. This highly undesirable result can be avoided by opting for the legal basis of Legitimate Interests for GDPR compliance purposes only, while separately complying with informed consent requirements for purposes of complying with clinical trial regulations (See EDPB Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection

Regulation (GDPR) Adopted on 23 January 2019 at https://edpb.europa.eu/our-work-tools/our-documents/opinion-art-70/opinion-32019-concerning-questions-and-answers-interplay_en). Rather than relying on GDPR-compliant consent in situations where the national requirements for clinical trial informed consent are too broad or vague because they were not designed under GDPR principles, the EDPB recommends the legal basis of Legitimate Interests for GDPR purposes. So long as national requirements for clinical trial informed consent are satisfied, Legitimate Interests may be used as a legal basis for GDPR data protection purposes which, combined with Article 9(2)(j) provided there is a national law in place (GDPR Article 9(2)(j) provides for the cumulative legal basis necessary to process special categories of data (such as medical data) for statistical, historical and scientific research) to avoid the undesirable consequences of the revocation of GDPR-based consent. However, as described above, one can only use Legitimate Interests for GDPR purposes if appropriate technical and organizational safeguards are implemented.

[xviii] The principle of functional separation exists under data protection regimes other than the GDPR using different terminology – e.g., heightened "De-Identification" under the California Consumer Privacy Act (CCPA) and the proposed Indian Data Privacy Law, and "Anonymisation" under the Brazilian Data Protection Law. The CCPA introduces the principle of functional separation through its definition of "Personal Information", which is subject to various protections under the Act. Personal Information includes "information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household." The CCPA's extensive list of data comprising protected Personal Information includes "static" and even "probabilistic" tokens (replacement identifiers) used to replace personal information if "more probable than not" that the information could be used to identify a consumer or device. While restrictions under the CCPA do not apply to "De-identified Data," traditional approaches to de-identification (e.g., HIPAA standards for de-identification) do not satisfy the heightened requirements for De-identification under the CCPA. CCPA heightened De-identification requirements are not satisfied using "static" and "probabilistic" tokens (replacement identifiers) because they fail to adequately separate information value from identity to prevent unauthorised reidentification of consumers.

[xix] "Personal information" under CCPA Section 1798.140(o)(1) means "information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household. Personal information includes, but is not limited to, the following if it identifies, relates to, describes, is capable of being associated with, or could be reasonably linked, directly or indirectly, with a particular consumer or household." CCPA Section 1798.140(o)(1)(A) provides that "identifiers include real name, alias, postal address, unique personal identifier, online identifier, internet protocol Internet Protocol address, email address, account name, social security number, driver's license number, passport number, or other similar identifiers." Under CCPA Section 1798.140(x), "Unique identifier" or "Unique personal identifier" means "a persistent identifier that can be used to recognize a consumer, a family, or a device that is linked to a consumer or family, over time and across different services, including, but not limited to, a device identifier; an Internet Protocol address; cookies, beacons, pixel tags, mobile ad identifiers, or similar technology; customer number, unique pseudonym, or user alias; telephone numbers, or other forms of persistent or probabilistic identifiers that can be used to identify a particular consumer or device." CCPA Section 1798.140(p) provides that a "Probabilistic identifier" means "the identification of a consumer or a device to a degree of certainty of more probable than not based on any categories of personal information." **Anonos heightened De-Identification capabilities enable the creation of dynamic de-identifiers which are not subject to the above CCPA definitions and therefore not subject to the restrictions of the CCPA.**

[xx] Consent exposes organizations to unpredictable interruptions in processing - e.g., when consent is revoked by a consumer directly or indirectly via its relationship with another party in the "data chain" - the obligation to delete the data flows with the data. Organizations can use Anonos' patented heightened de-identification-enabled Controlled Linkable Data to avoid these disruptions to operations.

[xxi] CCPA Section 1798.140(s)(2) provides that De-identification supports compliant "Research" using consumer information if (a) Pseudonymized or (b) aggregated.

[xxii] CCPA Section1798.145(a)(5) provides that De-identification supports lawful "Collection" of consumer information.

[xxiii] CCPA Section 1798.145(a)(5) provides that De-Identification supports lawful "Use" of consumer information.

[xxiv] CCPA Section 1798.145(a)(5) provides that De-Identification supports lawful "Retention" of consumer information.

[xxv] CCPA Section 1798.145(a)(5) provides that De-Identification supports lawful "Sale" of consumer information.

[xxvi] CCPA Section 1798.145(a)(5) provides that De-Identification supports lawful "Disclosure" of consumer information.

Gartner Cool Vendor

Why Cool: Anonos is cool because its patented technology creates relinkable non-identifying personalized data called **Variant Twins** that enable compliant analytics, ML, AI and data sharing.

ANONOS

LearnMore@anonos.com

IDC Report

Embedding Privacy and Trust into Data Analytics Through Pseudonymisation

Vendor Profile

Anonos: Embedding Privacy and Trust Into Data Analytics Through Pseudonymisation

Ralf Helkenberg

IDC OPINION

The digital transformation revolution is well underway. By leveraging data and technologies, organizations across all industry sectors are undergoing significant transformation in their business models. With greater demand for information, they are producing ever-more amounts of data. This data is being produced, processed, and shared in more places. Big Data adoption is set to rapidly increase beyond 2020 as more organizations use powerful analytics, increasingly infused with artificial intelligence (AI)/machine learning (ML) to achieve faster innovation, enhanced business performance, and competitive advantage. Cloud's promise of agility, scale, and flexibility is accelerating adoption. Organizations are embracing edge computing and hybrid multicloud architectures, shifting their compute strategies from centralized on-premises databases to distributed data infrastructure models. Distributed processing of large data sets and the richness of the data, much of it sensitive information, make Big Data highly open to data leakage and breaches. Meanwhile, privacy regulation is tightening, adding further complexities to turning data into business insight. In a GDPR and CCPA world, negligence of data privacy protections will not be tolerated and will result in higher fines. In this new reality, data privacy and security must follow the data.

Extract Value From Data in a Secure, Ethical Way

The sharing of data for the purposes of data analysis and research has many benefits. At the same time, concerns and controversies about data ownership and data privacy elicit significant debate. So how do organizations utilize data in a way that protects individual privacy but still ensures that the data is of sufficient granularity that analytics will be useful and meaningful?

IDC believes instilling trust in the use of data is a precondition for fully realizing the gains of data analytics. Encouraged through data protection regulation, encryption has become the default modus operandi for many to securing personal data. Yet it is primarily a security measure for making data unintelligible against unauthorized users. In environments where data is constantly moving between different parties and combined with other data, encryption – though providing effective data protection – is an inhibitor to creating valuable business insights. As a result, advanced analytics and data science projects that require fast access to data are slowing down or coming to a halt.

De-identification (functional separation) through pseudonymisation and anonymisation are important enablers of data analysis without requiring a compromise in data privacy and security. Although they may appear similar at first, they perform different functions in data protection law, such as the GDPR. The difference between anonymisation and pseudonymisation rests on whether the data can be re-identified. Data that has been irreversibly anonymised ceases to be personal data and does not require compliance with data protection law.

August 2020, IDC #EUR146734820

However, uncertainties exist as to whether such procedures can provide a sufficient degree of anonymity. Studies have shown that even within independent anonymised datasets, identifying individuals is not that difficult. Researchers were able to develop a machine learning model capable of correctly identifying 99.98% of Americans in any anonymised dataset using just 15 characteristics. A different MIT study of anonymised credit card data found that users could be identified 90% of the time using just four relatively vague points of information. This means knowing whether anonymisation has been achieved is rarely a black-and-white proposition and a challenging assessment to make. A further downside to anonymisation is that it results in a decrease in data utility. To preserve levels of utility, traditional anonymisation techniques restrict data processing to enclaves or silos to mitigate the risks of reidentification.

Pseudonymisation as a Way Forward

Pseudonymisation protects sensitive data by replacing one or more identifiers (direct or indirect) with pseudonyms or codes, which are kept separately and protected by technical and organizational measures. More importantly, the process can be reversed if the re-identification of data is required. While not a new privacy-preserving technique, pseudonymisation has been newly redefined and gained special prominence within the GDPR in which the benefits of proper pseudonymisation in protecting sensitive data are referenced 15 times. The European Union Agency for Cybersecurity (ENISA) provides further endorsement, where in its publication "Recommendations on shaping technology according to GDPR Provisions" highlights the following benefits from GDPR-compliant pseudonymisation:

- Pseudonymisation serves as a vehicle to "relax certain" data controller obligations, including:
  - Lawful repurposing (further processing) in compliance with purpose limitation principles
  - Archiving of data for statistical processing, public interest, scientific, or historical research
  - Reduced notification obligations in the event of a data breach
- Pseudonymisation supports a more favorable (broader) interpretation of data minimization.
- Pseudonymisation goes beyond protecting "real-world personal identities" by protecting indirect identifiers.
- Pseudonymisation provides for un-linkability between data and personal identity, furthering the fundamental data protection principles of necessity and data minimisation.
- Pseudonymisation decouples privacy and accuracy, enabling data protection by design and by default while enabling data about individuals to remain more accurate.

While pseudonymisation has many benefits, using it effectively requires significant expertise. The same ENISA report recognizes that effective pseudonymisation is highly context-dependent and requires a high level of competence to prevent compromise while maintaining data utility.

Data protection solutions that automate and simplify functional separation implementation have an important role to play in helping organizations realize their data strategies in a privacy-compliant manner. Technology vendors in this market space are few. One such company is Anonos, which have been at the forefront of shaping the pseudonymisation market with its BigPrivacy platform. The platform enables the sharing, collaboration, and analytics of personal data while enforcing security and data protection policies. It does this by employing state-of-the-art data de-identification technologies, controlled re-identification and re-linking of data, and a data-centric approach to security.

Its patented dynamic pseudonymisation technology differs from the more traditional approaches to data protection that use anonymisation, tokenization, static pseudonymisation, and generalization, and do not protect personal data from unauthorized re-identification when data sets are combined and used for multiple use. The BigPrivacy platform stands out for the range of pseudonymisation capabilities it can offer and its ability to meet specific technical requirements for achieving GDPR-compliant pseudonymisation. The flexibility in control settings that enable the relinking of de-identifiers back to individuals to support lawful business purposes is a plus.

A secure, scalable, and versatile platform, IDC believes BigPrivacy is a significant step up from traditional centralized data protection technologies, and it is well placed to resolve organizations' need for faster data insights while controlling data use risk across multiple environments and data-sharing partners.

IN THIS VENDOR PROFILE

This IDC Vendor Profile looks at Anonos, a technology provider of privacy-preserving data solutions. The profile focuses on the company's patented BigPrivacy platform, which pseudonymises personal data, thereby enabling organizations to undertake complex and sophisticated data analytics in a privacy- and security-compliant manner.

SITUATION OVERVIEW

Many organizations are trying to obtain more value from their data to improve their products and services. For example, more chief data officers and data analytical roles are being created to drive such data-enabled transitions. However, data privacy has become a flashpoint in the drive to achieve digital transformation. Concerns over potential privacy violations and the prioritization of locking data down through security measures has mistakenly led many organizations to forego the benefits of data insight. It need not be an either/or choice, since dynamic pseudonymisation organizations can achieve both.

Company Overview

Successful business partners for 20 years, Gary LaFever and Ted Myerson have a track record of developing privacy-preserving technology that help organizations turn regulation into a competitive advantage. Anonos was founded in 2012, on the premise that a whole new approach to data control, stewardship, and protection was necessary to unlock the maximum value of data and to turn it into a business asset without violating privacy, security, or regulatory restrictions. The outcome is the BigPrivacy pseudonymisation platform. The platform capitalized on growing market demand for privacy-compliance solutions, particularly the GDPR, and found adoption across financial, healthcare, telecom, and other data-intensive industries that rely on consumer data insights.

In 2019, the company secured a $12 million growth investment led by private equity firm Edison Partners.

Technology Proposition: BigPrivacy

Anonos' BigPrivacy platform supports a risk-based approach to data protection. This is accomplished by empowering data scientists and privacy engineers to set privacy controls at the data element level by applying a combination of traditional anonymisation techniques, GDPR-compliant pseudonymisation, and its own patented reidentification risk management technology – Controlled Linkable Data. Anonos enhanced the platform's pseudonymisation capabilities by leveraging the 50 technology recommendations by ENISA, and it can meet the specific technical requirements for achieving GDPR-compliant pseudonymisation.

The platform is comprises the following key components.

Variant Twins

The core of BigPrivacy's capabilities is centered around Variant Twins (i.e., the final pseudonymised dataset). The patented system leverages dynamic pseudonymisation to replace personal identifiers, such as a person's name and date of birth, with unique de-identifiers that prevent attribution of the data to a specific person without permission. Privacy control settings can be fined-tuned to provide the type and level of identifiability needed for each authorized use case. Because all Variant Twins are derived from the original source data, rather than permanently altered, organizations suffer no degradation in data value or accuracy.

GDPR-Compliant Pseudonymisation

Anonos' patented controlled relinkable dynamic de-identifiers are an advancement on the traditional pseudonymisation techniques of applying the same static tokens to direct identifiers across datasets. While useful in centralized environments, this traditional approach provides limited protection against unauthorized re-identification of individuals through data linkage and inference attacks. BigPrivacy uniquely combats the relinkage of data (Mosaic Effect) by using dynamic de-identifiers to introduce uncertainty (entropy) at the data element level for both direct and indirect identifiers. The product supports a range of pseudonymisation policies including the ENISA-defined fully randomized and deterministic pseudonymisation, and three additional intermediate-level policies: field, table, and document deterministic pseudonymisation.

Data Use Risk Management

K-anonymity sets out to address the risk of re-identification of anonymised data through linkage to other datasets – the higher the K value, the higher the degree of anonymity. The BigPrivacy platform incorporates a Data Use Risk Management module that leverages the k-anonymity concept. The pseudonymised dataset (Variant Twins) is filtered for reidentification risk to suppress records that do not meet the required k-anonymity threshold.

Lawful Insights API

Maximum data value often comes from combining and sharing data sets across multiple environments and with different partners. Anonos' Lawful Insights API enables lawful and multiparty processing both inside and outside of an organization's environment. Utilizing the same techniques and technology as BigPrivacy, it leverages endpoint interfaces to reduce data transfer friction and accelerate the process of safely and securely sending and receiving data for analytical processing. This is helpful when organizations experience trouble getting access to third-party data to augment the value of their data assets, or when third parties express concern about potential liability.

Company Strategy

Anonos' strategy aligns with addressing key privacy challenges in the sharing, collaboration, and analytics of personal data in a compliant manner to the GDPR. The main use cases are the following.

Compliant AI/ML Data Use

Data is a key driver for many of the new emerging technology innovations such as artificial intelligence and machine learning. Though AI/ML offer enormous business and innovation opportunities, they also pose privacy risks and regulatory challenges. The GDPR requires processing of personal data be carried out for specific purposes, no more personal data than is necessary to achieve those purposes is processed, and that personal data is only processed for as long as necessary to achieve those purposes. Tensions arise between these data privacy principles and AI, since the development of an AI system can often result in data being used for unexpected purposes, and often requires vast amounts of data to be inputted into the system for it to meaningfully detect patterns and trends. It also makes relying on consent as a lawful basis for many kinds of sophisticated data analysis impossible. BigPrivacy enables organizations to overcome the limitations of consent by using GDPR-compliant pseudonymisation to enable legitimate interests as a lawful basis for processing data.

IoT Data Protection

The number of devices connected to the Internet (i.e., Internet of Things or IoT) continues to grow exponentially. IDC forecast there will be 41.6 billion connected IoT devices generating 79.4ZB of data in 2025. Many of these devices exist within the automotive, healthcare, and consumer goods fields. Privacy and security though are big issues; through the data risk management controls of BigPrivacy, privacy-preserving data collection and management can be enabled across distributed IoT environments.

Compliant International Data Transfers

The GDPR sets out the legal mechanisms for the transfer of personal data outside the EU. The July 16, 2020, Schrems II decision of the European Court of Justice (CJEU) invalidated the EU-US Privacy Shield, a mechanism for transferring personal data from the EU to the U.S. At the same time, the CJEU reaffirmed the validity of standard contractual clauses, but added the caveat that they are only valid if they contain effective safeguards to ensure compliance with the protections provided by EU law. Anonos' BigPrivacy software is well placed to satisfy the Schrems II requirements for appropriate safeguards by creating pseudonymised versions of personal data (Variant Twins). Variant Twins ensure that desired processing results are achievable without providing third parties, including country authorities, the ability to re-identify individuals.

Data Protection by Design and by Default

The GDPR introduces the concept of "data protection by design and by default" into formal legislation, whereby organizations must integrate data protection into their processing activities and business practices from the design stage and throughout the life cycle. This includes adopting appropriate technical and organizational measures in implementing the data protection principles effectively. Pseudonymisation is one of several measures that organizations are urged to adopt to transition to the data protection by design and by default posture. The risk management controls in BigPrivacy help support data use minimization by enforcing selective access to data and ensuring employees only have access to the data required for them to do their jobs.

Data Processing in the Cloud

With the growing trend of moving workloads to the cloud, organizations can take advantage of cloud services without fear of breach of data privacy laws, as BigPrivacy pseudonymises the data at the point of ingestion. The solution enables organizations to create a global data lake in the cloud and also meet data sovereignty requirements.

FUTURE OUTLOOK

Concerns over data privacy and security have never been stronger, with the GDPR significantly influencing the way personal data is processed and protected by organizations. IDC believes many organizations have not yet fully recognized the benefits of GDPR-compliant pseudonymisation in deriving value from data while remaining compliant with data regulations and business rules. Many privacy professionals are not yet fully attuned to its potential, with their focus having concentrated in the past few years on ensuring regulatory compliance through policies and procedures. But a rapid shift is underway, with value-creation from data becoming a primary concern for organizations. In this new environment, privacy professionals are realizing they cannot just be compliance gatekeepers — they need to step up as business enablers to support organizations implement their data-driven business models. This means giving analytics and data science teams dynamic and frictionless access to datasets while enabling privacy-preserving measures to work in the background.

IDC thinks COVID-19 will be an inflection point for accelerated adoption of functional separation control techniques. Clinical and technological research projects that have arisen to mitigate the spread of COVID-19 have, in many cases, necessitated inter-organizational and cross-border data collaboration. Pseudonymisation has proved instrumental in providing a legally compliant approach to link and share sensitive datasets and provide access to secure analytical environments for researchers.

As the adoption of digital and data-driven business models accelerates and the need to use and share data in an ethical and trustworthy manner becomes a prerequisite, we believe Anonos is well-positioned with its state-of the-art pseudonymisation capabilities and granular protection control settings to capitalize on demand for privacy-preserving data analytics.

ESSENTIAL GUIDANCE

Advice for Anonos

Anonos has set itself up as an innovator and technology leader within the de-identification market space, and it is leading the charge to evangelize the benefits of pseudonymisation.

Though these concepts are not new, there remains much misunderstanding around the terminology and use of anonymisation and pseudonymisation. A legally and technically complex subject matter, Anonos needs to continue to push awareness around deployment best practices. Establishing a center of excellence in this field might further its cause.

It has rightly recognized the positioning of BigPrivacy as a legal compliance solution doesn't necessarily resonate with data audiences, but it is going in the right direction with its pitch that data privacy and data utility need not be an either-or choice and that both are possible with its leading-edge technology.

To prove its capabilities and credentials with new audiences, it needs to showcase through high-profile use cases how enterprises across different sectors have used its BigPrivacy platform to turn sensitive data into compliant business assets.

The technology also has a role to play in the hybrid-multicloud era. As distributed processing of Big Data and analytics migrates to cloud, cloud service providers are seeking to integrate and upgrade their privacy and security capabilities to secure this business, much of which is encryption focused. For Anonos, there is an opportunity to address this gap and enhance cloud providers' data application value proposition with its state-of-the-art pseudonymisation technology.

LEARN MORE

Related Research

- *IDC Market Analysis Perspective: European Data Privacy 2019* (IDC #EUR145752519, January 2020)
- *Anonos' SaveYourData - a EuroPrivacy Certified Solution "Deep Freezes" Enterprises' Existing Personal Data Sets as They Plan Analytics Strategies* (IDC #EMEA44411718, November 2018)
- *Europe's Political Leaders Put Ethics at the Heart of AI Strategy* (IDC #EMEA43324118, April 2018)
- *Anonos: Helping Businesses Become Data-Driven Without Compromising GDPR Compliance Obligations* (IDC #EMEA43641318, March 2018)

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

IDC U.K.

IDC UK
5th Floor, Ealing Cross,
85 Uxbridge Road
London
W5 5TH, United Kingdom
44.208.987.7100
Twitter: @IDC
idc-community.com
www.idc.com

Data Scientist Expert Opinion on Variant Twins and Machine Learning

ANONOS

DATA SCIENTIST EXPERT OPINION ON VARIANT TWINS AND MACHINE LEARNING

For Machine Learning tasks, Anonos Variant Twins provide performance comparable to clear text data. Results are virtually identical on every measure – with Variant Twins providing obviously enhanced resistance to re-identification.

Introduction

Mark Little, Chief Data Strategist, Anonos

The public's interest in the reasonable expectation of privacy is met when personal data remains private: within reasonable limits, the data cannot be used to single out, or to inferentially identify or link personal data to a particular data subject. Historically, privacy has been maintained by reducing access to identifiable data, while ensuring that the likelihood of re-identification, largely interpreted through equivalence classes, is reduced. However, personal data is increasingly being collected outside of traditional situations, contains increasingly detailed information, and is being utilized by new entities for new purposes. Yet, the public maintains the same interest in privacy.[1]

One long-used method in de-identification has been to consistently substitute the same token for a given direct identifier each time it occurs. While perhaps effective decades ago, this approach has long been known to no longer provide meaningful protection.[2] Historically this technique has been referred to as pseudonymisation though the term did not have the benefit of any statutory underpinning.

That changed with the arrival of the EU General Data Protection Regulation (GDPR). While many are aware that the GDPR for the first time established a formal regulatory definition for Pseudonymisation, few are aware of how significantly that definition changes what must be done to advance a claim that a data set has been Pseudonymised in compliance with the GDPR. To fully appreciate the magnitude of the changes, consider the following:

- The GDPR defines pseudonymisation as (emphasis added):

  > " the processing of **personal data** in such a manner that **the personal data** can **no longer be attributed** to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;"[3]

- While the scope for pseudonymisation has been historically limited to direct identifiers, or what is often referred to as personally identifying information (PII, such as names, ID numbers, email addresses, etc.), under the GDPR it encompasses Personal Data, which is defined as (emphasis added):

  > "…**any information relating** to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who **can be identified**, directly or **indirectly**, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to **one or more factors specific to** the physical, physiological, genetic, mental, economic, cultural or social identity of **that natural person**…"[4]

---

[1] Clouston, Sean, Data Privacy in An Age of Increasingly Specific and Publicly Available Data: An Analysis of Risk Resulting from Data Treated Using Anonos' Bigprivacy Methodology, Appendix 1 – Data Scientist Expert Opinion on BigPrivacy, p.3, Blueprint for GDPR, 2nd Edition, Jan 2019, p. 48.

[2] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization (August 13, 2009). UCLA Law Review, Vol. 57, p. 1701, 2010, U of Colorado Law Legal Studies Research Paper No. 9-12, Available at SSRN: https://ssrn.com/abstract=1450006

[3] GDPR Article 4(5), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&rid=3

[4] GDPR Article 4(1),https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&rid=3

- Read together the implications are clear:

    1) Pseudonymisation is an **outcome** (not simply a technique) that describes a **dataset** (“any information”) containing personal data, not just individual fields within a dataset.

    2) If an unprotected data set contains direct identifiers, or information that could be used to **indirectly identify** a data subject, **the entire data set is personal data.**

    3) It is **NOT possible to achieve Pseudonymisation without protecting so-called indirect or quasi-identifiers**, because “identifiable” includes the ability to use combinations of factors to indirectly identify a data subject,

    4) It is also **NOT possible to conclusively demonstrate Pseudonymisation has been achieved if static tokens that are consistent across datasets are used**, due to the possibility of attributing personal data to a specific data subject by combining datasets.

As described in detail above in the main body of this document, Anonos technology uses patented technology to achieve GDPR pseudonymisation by enabling the assignment of dynamic pseudonyms to both direct and indirect identifiers. But it goes further, by combining this approach with traditional anonymisation techniques along with patented enhancements to pseudonymisation to enable Controlled Relinkable Data that ensures no loss in the utility of source data--which by definition anonymous data cannot do--while providing resistance to re-identification that is comparable or superior to other approaches. The data assets that deliver this powerful combination are called Variant Twins.

What follows is an analysis of the comparative utility of Variant Twins and the corresponding original clear text data sources for Machine Learning and AI.

ANONOS

Analysis: "The Comparative Utility of Clear Text and Variant Twins in Machine Learning and Artificial Intelligence Model Development"

Mike Nemke, Director of AI & Machine Learning, Aptive Resources

Mike is the Director of AI & Machine Learning at Aptive Resources where he leads the rapidly expanding Data Science Practice and ML Product development efforts for several large government clients. Prior to his time with Aptive Resources, Mike spent 5 years working as a Lead Data Scientist for startups and consulting firms in the SF Bay Area. Mike also has an MS in Data Science from Northwestern University.

Motivation

Data science and machine learning practices are emerging and operating in a complex space, and Data Science professionals are subject to the same inherent systemic and cognitive biases[5] as any other role. Unfortunately, due to the increasingly automated and predictive nature of data work, those otherwise marginal biases, conscious or otherwise, and their effects, are magnified. Additionally, these data sets frequently contain personal data that is increasingly subject to restrictions in processing under emerging data protection and privacy regulations. In order to minimize possibly harmful bias and comply with data privacy and protection requirements for work produced by myself and my team of Data Scientists, Data Analysts, and Machine Learning Engineers, I decided to test options for masking or pseudonymisation of the data we work with.

Situation

Given an artificially generated dataset with 100,000 job applicants, we want to predict, without exposing their personal information, whether they will accept or decline a job offer. Given the significant amount of personal information in the form of direct and indirect identifiers, and a desire to remove systemic biases from hiring practices, we want to compare analysis of the clear text dataset against a pseudonymised dataset (i.e., an Anonos Variant Twin) to assess the relative utility given the significantly enhanced resistance to unauthorized re-identification of the Variant Twin.

Analysis Goal

Analyze and measure the performance of analytical models on the clear text dataset and the Variant Twin. The analysis performed is a series of classification models to identify whether candidates will accept or decline a job offer.

Intro to the data

The original clear text data consisted of 100,000 fictitious job applicants and included name, email, age, gender, race, location, education, work experience, as well as application and interview specific data. The Variant Twin dataset used the original data, but with pseudonymised replacements for gender, race, state, degree, college, college major, job titles, job tenures, job applied for, and application source. The name, email, city, and interviewer note data were removed in the Variant Twin generation process and age remained unchanged. Non identifiable attribute information (e.g., interview ratings) also remained unchanged.

[5] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. 2020. https://arxiv.org/pdf/1811.07867.pdf.

ANONOS

Clear Text Data

	ID	First_Name	Last_Name	Email	Gender	Race	Age	City	State	Highest_Degree	College
0	13462002	Napoleon	Heger	Napoleon.Heger@teleworm.us	male	White American	23	Saint Louis	MO	BS	UT Austin
1	24859940	Pauline	Henderson	Pauline.Henderson@superrito.com	female	White American	25	Portland	OR	MS	Michigan State University
2	34297375	Larry	Hundley	Larry.Hundley@fleckens.hu	male	Black or African American	26	Honolulu	HI	MS	UT Austin
3	93713586	Michael	Miron	Michael.Miron@jourrapide.com	male	Asian American	28	Pittsburgh	PA	BS	Michigan State University
4	39059207	John	Austin	John.Austin@armyspy.com	male	White American	26	Saginaw	MI	MS	Florida International University

Variant Twin dataset

	row_rddid	Gender	Race	Age	State	Highest_Degree	College	Major	Job1_Title	Job1_Tenure	Job2_Title	Job2_Tenure	Job3_Title
0	RD-0cd7ba3b-a18a-4ebf-9c6b-8ea0e595e2a0	gender-1904	race-5a85	23	state-e938	degree-a67d	college-02da	major-942e	job1-b9f3	3.0	job2-5890	NaN	job3-5890
1	RD-acc741f2-0ec0-47e8-a859-47604f42e1a4	gender-9b08	race-5a85	25	state-3af3	degree-9c23	college-c75d	major-e91b	job1-b884	4.6	job2-5890	NaN	job3-5890
2	RD-2fd28190-cd0a-45b9-868b-dc1f3b053f90	gender-1904	race-d7e7	26	state-4833	degree-9c23	college-02da	major-0c06	job1-ff06	1.4	job2-5890	NaN	job3-5890
3	RD-ae325644-47bb-4c00-be26-4529cf22ad2e	gender-1904	race-7c69	28	state-0de7	degree-a67d	college-c75d	major-942e	job1-e2bd	7.0	job2-5890	NaN	job3-5890
4	RD-5df6020a-c3ff-4056-9eae-eacf19d5c630	gender-1904	race-5a85	26	state-3f7b	degree-9c23	college-ced7	major-dc0f	job1-b9f3	2.0	job2-5a15	1.4	job3-b884

Methodology

Tools used

All of the code for this analysis was completed in python using Jupyter notebooks. The packages used for this analysis include Matplotlib and Seaborn for data visualizations, NumPy and Pandas for shaping and managing the data, and TensorFlow and Scikit-learn for modeling and analyzing the data.

Data Preparation

The target variable for this analysis is the ‘Decision’ variable. The Decision variable was either ‘Accepted’ or ‘Declined’ in both datasets and was converted to a binary variable with ‘0’ mapping to ‘Declined’ and ‘1’ mapping to ‘Accepted’. All candidates that did not receive an offer, and therefore had a null decision value, were dropped from the analysis. Both datasets contained 28,054 instances with either a ‘0’ or ‘1’ decision value.

ANONOS

Due to null values in the unmasked dataset, the following columns were dropped from both datasets: Terminated, Job2_Title, Job2_Tenure, Job3_Title, Job3_Tenure, Offer. The 'Terminated' variable indicated an employee that accepted an offer and had later terminated employment. That means the variable is only viable for the candidates that accepted an offer. The job title and tenure columns were largely null. Finally, the 'Offer' variable was 'yes' for any candidate who would have 'Accepted' or 'Declined' an offer, thereby providing no analytical value in this exercise.

In order to use the categorical variables in the datasets to predict whether a candidate would accept or decline an offer, the following variables were encoded as categorical variables using Pandas get_dummies() function: Gender, Race, State, Highest_Degree, College, Major, Job1_Title, Job_Applied_For, and Source.

The input variables used in this analysis were: Gender, Race, Age, State, Highest_Degree, College, Major, Job1_Title, Job1_Tenure, Job_Applied_For, Source, Assessment_Score, Interviews, Interview_Score_Avg, Recruiter_App_Eval, Cycle_Time_Days. Following variable encoding, there were a total of 161 input variables.

Modelling Technique 1: Random Forest

The datasets were imbalanced with 19,660 offers accepted and 8,394 offers declined. Due to the imbalance, the first classification technique used was a Random Forest Classifier using a Scikit-learn module (link). Three separate approaches were taken to running the Random Forest Classifiers, first using no hyperparameter tuning, the second using class_weight='balanced', and finally class_weight='balanced_subsample'. For this hyperparameter:

- The "balanced" mode uses the values of the target variable (Decision in this case) to automatically adjust weights inversely proportional to class frequencies in the input data as `n_samples / (n_classes * np.bincount(y))`
- The "balanced_subsample" mode is the same as "balanced" except that weights are computed based on the bootstrap sample for every tree grown.

Clear Text Random Forest Classifier:

```
rf = RandomForestClassifier(random_state=43)
rf.fit(X_train_clear, y_train_clear)
y_pred_clear_rf = rf.predict(X_test_clear)
print(classification_report(y_test_clear, y_pred_clear_rf))
```

```
              precision    recall  f1-score   support

         0.0       0.28      0.15      0.20      1627
         1.0       0.71      0.84      0.77      3984

    accuracy                           0.64      5611
   macro avg       0.49      0.50      0.48      5611
weighted avg       0.58      0.64      0.60      5611
```

ANONOS

Variant Twin Random Forest Classifier:

```
rf = RandomForestClassifier(random_state=43)
rf.fit(X_train_variant, y_train_variant)
y_pred_variant_rf = rf.predict(X_test_variant)
print(classification_report(y_test_variant, y_pred_variant_rf))
```

```
              precision    recall  f1-score   support

         0.0       0.29      0.16      0.21      1627
         1.0       0.71      0.84      0.77      3984

    accuracy                           0.64      5611
   macro avg       0.50      0.50      0.49      5611
weighted avg       0.59      0.64      0.61      5611
```

As you can see from the figures above, precision, recall, and f1-scores are all virtually the same for models with the class_weight hyperparameter omitted. The same outcome resulted for both the balanced and balanced_subsample runs.

Modeling Technique 2: Neural Network

In preparation for the neural network, both datasets were converted to arrays and scaled using the Scikit-learn standard scaler module (link). The model was developed using TensorFlow and Keras. 3 dense layers, and 2 dropout layers made up the model, with the final dense layer using a sigmoid activation function. The hyperparameters used for the NN: Adam optimizer with a learning rate of 1e-3, and binary cross entropy as the loss function. For each dataset, the model was set to run 100 epochs with a batch size of 500.

Clear Text NN Model:

```
model_clear = make_model(train_features_clear)
model_clear.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 32)                5184
_________________________________________________________________
dropout (Dropout)            (None, 32)                0
_________________________________________________________________
dense_1 (Dense)              (None, 32)                1056
_________________________________________________________________
dropout_1 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 33
=================================================================
Total params: 6,273
Trainable params: 6,273
Non-trainable params: 0
_________________________________________________________________
```

ANONOS

Clear Text Neural Network Classifier Outcomes:

```
loss :  0.6355025006517585
tp :  3734.0
fp :  1613.0
tn :  71.0
fn :  193.0
accuracy :  0.67813224
precision :  0.6983355
recall :  0.95085305
auc :  0.49697286

Legitimate Declines (True Negatives):  71
Decline Incorrectly Labeled as Accepted (False Positives):  1613
Accepted Incorrectly Labeled as Declined (False Negatives):  193
Legitimate Accepted (True Positives):  3734
Total Offers:  3927
```

Variant Twin Neural Network Classifier Outcomes:

```
loss :  0.6419994837565125
tp :  3475.0
fp :  1451.0
tn :  208.0
fn :  477.0
accuracy :  0.65638924
precision :  0.7054405
recall :  0.8793016
auc :  0.50049025

Legitimate Declines (True Negatives):  208
Decline Incorrectly Labeled as Accepted (False Positives):  1451
Accepted Incorrectly Labeled as Declined (False Negatives):  477
Legitimate Accepted (True Positives):  3475
Total Offers:  3952
```

There are inherent variations in the outcomes due to the use of neural networks, but as shown in the figures above, the Variant Twin returned results very comparable to clear text, underperforming modestly in predicting the majority class (“Accepted’ offers) and slightly outperforming in predicting the minority class (‘Declined’ offers).

Modeling Technique 3: Logistic Regression

Using the Scikit-learn logistic regression module, both datasets were used to fit and predict the decision variable.

Clear Text Logistic Regression Outcomes:

```
print(classification_report(y_test_clear, y_pred_clear))
```

	precision	recall	f1-score	support
0.0	0.28	0.46	0.35	1627
1.0	0.70	0.52	0.60	3984
accuracy			0.50	5611
macro avg	0.49	0.49	0.47	5611
weighted avg	0.58	0.50	0.52	5611

Variant Twin Logistic Regression Outcomes:

```
print(classification_report(y_test_variant, y_pred_variant))
```

	precision	recall	f1-score	support
0.0	0.28	0.46	0.35	1627
1.0	0.70	0.52	0.60	3984
accuracy			0.50	5611
macro avg	0.49	0.49	0.47	5611
weighted avg	0.58	0.50	0.52	5611

As can be seen in the figures above, the outcomes were identical.

Discussion of Results

Consistent results across both datasets
In this analysis, we used random forests, neural networks, and logistic regression to classify whether a job candidate would accept an offer with 2 datasets. One dataset was the original clear text data and the other was an Anonos Variant Twin. The analysis shows that the Anonos Variant Twin performed virtually the same as the clear text data from which it was derived for machine learning tasks. The differences in model performance were negligible and the only real variance came from inherently variable models (neural networks).

Conclusions
For Machine Learning tasks, Anonos Variant Twins provide performance comparable to clear text data. Both datasets were virtually identical on every measure – with Variant Twins providing obviously enhanced resistance to re-identification.

In my work with Aptive Resources as Director of AI & Machine Learning, I commonly lead the development of data science and machine learning models and products for large U.S. government clients. Our clients prioritize privacy of personal data in all of our projects and based on my experience with competitive products and approaches, Anonos' Variant Twins approach is best in class. Based on our experience with this testing, we plan on adopting Anonos technology for analytics projects with datasets rich in personal data.

Data Scientist Expert Opinion on BigPrivacy

TITLE: DATA PRIVACY IN AN AGE OF INCREASINGLY SPECIFIC AND PUBLICLY AVAILABLE DATA: AN ANALYSIS OF RISK RESULTING FROM DATA TREATED USING ANONOS' BIGPRIVACY METHODOLOGY.

December 2, 2017

Author: Sean Clouston, PhD, #3-096, Stony Brook University, Health Sciences Center, 101 Nicholls Rd., Stony Brook, NY, 11794.

Conflicts of interest: The original version of this report was started by Dr. Clouston when he was an external and independent consultant for Anonos, at which time the author had no conflicts of interest to report. However, during the process of the finalization of the initial, and this updated version of the report, but before publication of the initial version of this report, the author became an Anonos shareholder. Anonos played no role in the study conclusions, and did not direct the analysis. Anonos provided editorial suggestions relating to the implications and implementations of the analytic results provided herein.

Note: Changes from the original 2015 report, as reflected in this updated December 2, 2017 report, are limited to clarifications that the efficacy of the BigPrivacy methodology applies equally in situations where an equivalence class is comprised of five members.

ANONOS®

Abstract

The public's interest in the reasonable expectation of privacy is met when personal data (PD) remains private: within reasonable limits, the data cannot be used to single out, or to inferentially identify or link PD to a particular data subject. Historically, privacy has been maintained by reducing access to identifiable data, while ensuring that the likelihood of re-identification, largely interpreted through equivalence classes, is reduced. However, PD is increasingly measured outside of traditional situations, contains increasingly detailed information, and is being utilized by new entities in new purposes. Yet, the public maintains the same interest in privacy. As a result, static de-identification tactics, which delete known identifiers by replacing them with one token used consistently for an individual, have been increasingly susceptible to critique.

One way forward is to use new temporally dynamic obscurity protocols that actively minimize the risk of re-identification. This report analyzes the ability of BigPrivacy to minimize re-identification under different circumstances and thus to ensure data privacy. Analyses provided in this report aid in assessing privacy and security risks and maintaining privacy and security when data incorporates detailed and even longitudinal information. PD is kept private within an acceptable level of risk, subject to some constraints on oversight and sample size. Data stored or transmitted pursuant to these methods is de-identified in the traditional sense; further, data storage and transmission are more robust using these methods for a number of reasons outlined in the report. Moreover, because data security is dynamic, privacy policies can be flexibly implemented to ensure security is consistently and completely ensured.

ANONOS®

Introduction

Data privacy is important for many reasons. As just one example, health data, once released, could be used to reveal realities about a person's health that he or she may not want to share, including diagnoses of societally stigmatized diseases (e.g., PTSD, HIV/AIDS, Schizophrenia, etc.) and health issues having financial implications for their families or their health insurance carriers (e.g., physical activity, blood pressure, financially discriminatory actions, etc.). On the other hand, as Matthews and Harel (2011) highlight in their review of data privacy issues and solutions, researchers must ensure that data are accessible for potentially valuable research applications. Moreover, data are increasingly collected as by-products of private sector innovation, but while these data need to be protected, it is not in the public interest to stifle innovation requiring that data. Static de-identification, which we here define as de-identification that maintains the structure and linkages of data and achieves de-identification by replacing identification data with a randomized static token, was previously deemed to be sufficient for reducing the risk of re-identification. However, new data, more powerful types of data analytics, and the increasing number of data sources have made researchers, policymakers, and software developers sceptical this can continue (Chen & Zhao, 2012; de Montjoye, Radaelli, Singh, & Pentland, 2015).

As one example of how regulations are affected by the issues surrounding data minimisation, a U.S. Federal Trade Commission report noted that while HIPAA traditionally only pertains to a small number of people handling health information, such as doctors or hospitals, "health apps are [increasingly] collecting this same information through consumer-facing products, to which HIPAA protections do not apply…" and goes on to state that "consumers should have transparency and choices over their sensitive health information, regardless of who collects it" (Federal Trade Commission, 2015). The conclusion of the FTC report was twofold: the majority decision supports the need for "data collection minimisation," or the wholesale deletion of collected information from the information ecosystem, while the minority decision notes that this form of data minimisation might negatively impact health-related research and decision making (Federal Trade Commission, 2015).

The FTC dissent highlights the contrast between data value and data privacy. A non-partisan research firm (the Information Technology and Innovation Foundation or ITIF), further highlights problems with data collection minimisation in the private sector: "the FTC's report correctly recognizes that the Internet of Things offers potentially revolutionary benefits for consumers and that the industry is still at an early stage, [but the report] unfortunately attempts to shoehorn old ideas on new technology by calling for broad-based privacy legislation"; further, "in calling for companies to reduce their use of data, the FTC misses the point that data is the driving force behind innovation in today's information economy" (Castro, 2015). These dissenters each view such data collection and analysis efforts as *serving the individual and public interests*, even at the cost of privacy. *Wired* magazine concretizes these dissents, reporting that though IoT devices currently being developed are geared towards gathering "reams of largely superficial information for young people whose health isn't in question, or at risk" (Herz, 2014), "the people who could most benefit from this technology—the old, the chronically ill, the poor—are being ignored… [primarily because] companies seem more interested in helping the affluent and tech-savvy sculpt their abs and run 5Ks than navigating the labyrinthine world of... HIPAA."

ANONOS®

Three Main Limitations

There are three main limitations with the current approach to data privacy. First, static de-identification is not robust. Second, transmission is particularly problematic. Third, an increasing number of entities are involved in providing guidance about privacy in a way that is increasingly difficult to maintain. These are discussed in detail in the following section.

First, data are not truly de-identifiable. Specifically, a recent article in *Science* observed the relative ease with which one can uniquely identify individuals using only small amounts of financial information (de Montjoye et al., 2015): indeed, for 90% of the cited sample only four pieces of information were needed to achieve "unicity" – i.e., development of unique identifying profiles derived from traditionally de-identified financial data. As noted by (El Emam, 2015), unicity in the dataset does not mean that any person has successfully re-identified each individual; however, once de-identified and made available to the public, data are subject to "data fusion", which is the linking of multiple different datasets together in order to broaden our understanding of the people in a dataset. The risk of data fusion has led to this finding being highly publicized, for example the *Wall Street Journal* noted that unicity in financial data meant one could readily "find the name of the person in question by matching their activity against other publicly available information such as LinkedIn and Facebook, Twitter, and social-media check-in apps such as Foursquare" (Hotz, 2015). The *Harvard Business Review* concludes the implications of this work "are profound. Broadly, it means that static anonymity doesn't ensure privacy" (Berinato, 2015).

Second, current data de-identification tactics when data are being transmitted are especially questionable, such transmission increasingly occurring through devices considered to be within the "Internet of Things" (IoT) (Rivera & van der Meulen, 2014). During transmission, the normal "size" of the dataset is curtailed, further weakening the assumptions on which we rely for data security. At the same time, the amount and specificity of data is increasing with data available such as the person's everyday progression from home to work or the number of calories, types of food, and restaurants that they ate in during their last week. Furthermore, IoT devices increase the importance of information transmission; for example, in the case of healthcare information, clinicians might be able to use interconnected devices to monitor a patient's health, including vital signs or physical activity, potentially raising new concerns regarding data privacy and an increased risk of data breach (Tyrrell, 2014).

Finally, de-identification, proceeding in a static manner, must be implemented under one specific policy regime to the detriment of others. For example, it may be that data collected under one policy are made more or less secure than are necessary under a newer or different structure so that data managers either must redo their work to a different standard, resulting in substantial inefficiency, or may simply choose not to allow access to data because the cost is too large to ensure compliance. In such a circumstance, having pre-approved levels of access could help to ensure that data are both accessible to individuals from varying regions or policy regimes, and at varying levels of security.

ANONOS®

Data Construction

Without a solution that responds to these concerns and truly de-identifies PD, a broad range of individuals including, but not limited to, software developers and information technology specialists, will have access to non- de-identified PD data. Static de-identification, as noted above, is not working. Dynamically obscuring data may be one way to retain data privacy while reducing the risk involved in collecting, storing, and analysing such data (Warren, 2014). Examining new ways requires a basic knowledge of dataset construction, different types of data, and existing de-identification protocols. The following sections provide an introduction to topics underlying these issues before moving on to a more formal analysis of the risk of re-identification.

De-identification

The process of de-identification decreases privacy risks to individuals by removing identifying information from protected or personal data (PD). Thus, in the dataset presented in Table 1 below, data from the first column (identifying information, here an IP address) would need to be removed from the dataset. We should note, however, that de-identification usually references two somewhat separate processes: the removal of the certainty that any particular individual is part of the observed dataset, and the removal of the certainty that any particular observation might, in the correct circumstances, be sufficiently unique to be re-identified with other available
data. Thus, while it is often believed that removing these IP addresses renders similar datasets (usually with more observations) “de-identified” in the traditional sense, as discussed above many of these observations can be uniquely identified using data characteristics that can lead to “unicity” within the database, rendering them unique in the data and thereby at risk of re-identification (de Montjoye et al., 2015).

The problem of de-identification has historically been addressed in a number of temporally static ways. Matthews and Harel (2011) list the following techniques used for de-identification: 1) limitation of detail, 2) top/bottom coding, 3) suppression, 4) rounding, 5) adding noise, and 6) sampling. *Limitation of detail* works through categorizing or collapsing information to reduce the possibility of characteristic re-identification. *Top/bottom coding* characterizes the replacement of observational data with a “top-code”, an upper limit on all published values of a variable, and/ or a “bottom-code”, a lower limit on all published values of a variable, the replacement of which reduces the likelihood of re-identification of data that are more likely to be unique, such as very high incomes, by recoding them so that outlying observations are grouped together. *Suppression* works by removing potentially identifiable data from the publicly available dataset. *Rounding* introduces noise by randomly re-assigning rounded values to all the individuals in a dataset, and is often used for ages because though we may know that individuals are 45.67 years old (i.e., 45 years and 8 months), we recode that information into yearly (as age 45) or into age groupings (such as 45-49). *Addition of noise* uses a randomization routine to change the values in each cell by some random amount, an approach often used with geographic data such as that in the *Demographic and Health Surveys*, which have randomly dispersed geographic residential locations by some distance less than five kilometers (Measure DHS & ICF International, 2013). Finally, *sampling* resolves de-identification by requiring that data released be only a subset of the data available, with the convention that between 95-97% of the data collected could be released; however, sampling also resolves another issue, notably that individuals should not be known to have been a part of the dataset, because it removes, at random, entire individuals from a dataset so that you may not be certain that any particular person who was originally contained within the dataset are also contained within the dataset released.

Since then, more complex mathematical routines have been used to ensure data is kept confidential and that this confidentiality is unlikely to be broken. The most useful of these build on the randomization approach because it is the most secure and removes the least value from the data. *Matrix masking*, for example, codifies the data by multiplying them by a form of encryption key that researchers must know about and account for when analysing data (Cox, 1994). Another particularly interesting method, called *synthetic data,* uses data matching methods originally built to provide results for missing data to effectively swap characteristics between individuals in a dataset, thereby retaining some of the statistical uniqueness and value while reducing the risk of re-identification (Rubin, 1993). Note that these methods work by *increasing uncertainty* about the identity of any one individual without unnecessarily modifying the data itself.

De-identification

Current static de-identification methods generally rely on data being purposefully collected into one dataset before being anonymized, using usually one or two of the previously mentioned techniques, to protect privacy. These datasets are then shared with the public either at large or through limited access policies depending, largely, on the level of specificity provided in the data, or kept hidden from the public entirely.

This data fusion may be increasingly easy to do, especially as individuals “check in” on publicly available social networking sites. These data fusion techniques are big, flexible, fast, constantly changing, and being made more specific; they also use data that are being increasingly used, held, or transmitted by an ever larger number of individuals. Thus, the conclusions of de Montjoye et al. (2013, 2015) are likely true: if data are unique, those data may be readily identifiable among a substantial portion of users.

Estimating risk of re-identification

The probability of re-identification can be estimated (El Emam, Dankar, Vaillancourt, Roffey, & Lysyk, 2009). Re-identification requires access to the full datasets, which contain information for five separate individuals, and that re-identification is being undertaken purposefully to find a particular individual who may be in the dataset. We provide an example of a blood-pressure monitoring application on an internet-connected wearable device (i.e., a phone, watch, shoe, or other wearable device), which 1) identifies a user, 2) monitors health information, and 3) is linked to geographic positioning systems (GPS) data. Because this is a useful and clearly risky subject, we will rely on this example throughout this analysis. To protect the privacy of individuals associated with these data, we are faced with the following difficulties: de-identification may be subject to difficulties in both data storage and in data transmission, and further may propose a risk to multiple governing bodies since it collects data that may include health information, could easily integrate different types of data including longitudinal data, and may also have clinical applications if clinicians are interested in monitoring patients’ everyday health in this way.

Below we define a data matrix or dataset ***H*** (Table 1), which for simplicity is a 5 x 6 matrix that contains 30 unique data points (called cells). Different rows may contain different information for the same individual if that person is followed over time or is observed by different people (in longitudinal or higher-dimensional data). Note that there are therefore both explicit and implicit identifiers within most datasets: the IP address is explicit while the row number, when not longitudinal data, is an implicit identifier.

ANONOS®

Table 1. Hypothetical dataset (**H**) collected from multiple smartphones on the same network by a blood pressure application in December 2014

IP address	Latitude	Longitude	Age	Sex	High blood pressure
192.168.0.0	40.13	-79.85	32	M	No
192.168.101.201	40.13	-79.86	45	F	Yes
192.168.4.57	40.15	-79.54	39	M	No
192.168.22.40	40.29	-79.54	56	M	No
192.168.1.220	40.29	-79.56	42	F	No

Note: IP addresses are assumed static for period of observation; Latitude and Longitude are "rounded off" to two decimal places and so are not precise to the level of a specific house.

Using the example in Table 1 above, we will suppose that *Alice* is the person with high blood pressure corresponding to IP address 192.168.101.201. Let us assume we know the type of information a neighbour, co-worker, or employer might know about *Alice*. Suppose, for example, we know she is female, that she was born approximately 40-50 years before the data were collected, that we know that she lives in Belle Vernon, Pennsylvania (Latitude, Longitude = +40.13, -79.86). However, we want to know further whether *Alice* has high blood pressure, and thus we also need to re-identify her using the data provided.

We follow previous reviews of re-identification risk assessment (El Emam et al., 2009) that define an "acceptable risk" as one that is at most = 0.20; further, for reasons that will become clear later, we further clarify that an "unacceptable risk" is one known to be greater than = 0.20. Then, the whole dataset's risk of re-identification (*r*) can be defined as: , where *f* is the number of individuals with equivalent characteristics to any one particular person, including here *Alice*, and min_j is the minimum number of individuals in a subset of categories (*j;* sometimes called an equivalence class, the basis for the "unicity" argument) that fit *Alice*'s known characteristics.

The risk has also been specified in the following way, including the measurement of how many equivalence classes that there are in a dataset who are expected to be distinct. Along those lines, Benitez and Malin (2010) provide the following definition of total risk: where *k* references the number of individuals in *b* possible equivalence classes for a sample of size *n*. Because the total risk revolves around the risk within a particular equivalence class, we thus begin by briefly overviewing equivalence classes in re-identification, before explaining why data parsing helps secure privacy.

ANONOS®

Re-identification in practice

In table 1 above, our first efforts would be to rely on Anonos BigPrivacy to mask and eliminate IP addresses, replacing that information with Anonos BigPrivacy patented dynamically pseudonymised tokens (referred to herein as “Dynamic De-IDentifiers” or “DDIDs”). Prior to this, the risk of identification is perfect: $r = 1/1 = 1$. This is an unacceptable risk because $r = 1 > = 0.20$. However, after removing IP addresses, the risk is reduced because we cannot rely on identifiable information. In particular, the risk becomes $r = 1/n = 1/5 = 0.20$, a borderline but acceptable risk. We still, however, know that *Alice* is a woman aged 40-50 who lives in Belle Vernon, Pennsylvania. The risk of re-identification as a woman is defined as the inverse of the number of people in the equivalence class: in this case, the inverse of the number of women in the dataset, and thus . Because 0.5 is larger than 0.2 (as defined above by our categorization of acceptability,), we note that this is already an unacceptable level of risk. However, for clarification as to the nature of data linkages we push this data further to use more characteristic and specific data. We examine the data and note that there are two women aged 40-50 in the data. Therefore, we calculate r = ½ = 0.50; since r > = 0.20 this remains an unacceptable risk. We further know that *Alice* lives in Belle Vernon (latitude ranges from 40.12 to 40.14, longitude ranging from -79.84 to -79.86). This shows us that there are two people in these data living in Belle Vernon, and thus we calculate r = ½ = 0.50; since r > = 0.20 we define this as an unacceptable risk. Linked together, we can further see that, of those people living in Belle Vernon, only one is a female aged 40-50. Thus, data linking increases our risk of re-identification to $r = 1$, an unacceptable risk resulting in certain re-identification.

Table 2. Hypothetical BigPrivacy data (**H**) collected from multiple smartphones on the same network by a blood pressure application in December 2014

DDID	IP address	DDID	Lat.	DDID	Long.	DDID	Age	DDID	Sex	DDID	High BP
5657	192.168.4.57	5934	40.13	5049	-79.86	4958	42	5141	F	6878	No
5854	192.168.101.201	3030	40.29	3060	-79.85	3938	32	6236	M	4948	No
3938	192.168.0.0	1234	40.13	9090	-79.54	5010	45	7747	M	4094	No
5910	192.168.1.220	1410	40.15	8974	-79.54	7079	56	8585	M	0967	No
2039	192.168.22.40	4040	40.29	9030	-79.56	7078	39	9999	F	0847	Yes

Note: We have shortened title names to limit table size. Lat.: Latitude; Long.: Longitude; BP: Blood Pressure; DDID: Dynamic de-identifiers. Each DDID references the value in the column to its right. For ease of reference, the above table is represented to include both DDIDs and corresponding data; in an actual implementation of the BigPrivacy method, this table would not contain BigPrivacy keys that would be stored in a highly secure master look-up database which contains information necessary to, under technically controlled conditions, relink all connections between direct and indirect (structured and unstructured) identifiers and DDIDs.

Removing information from the implicit linkages is novel because it removes the possibility of linking data together to contextualize information. Thus, data are not saved in Table 1 above, but in a way that more closely resembles Table 2 above. For ease, the following discussions reference the "worst case scenario," wherein an entire dataset with small sample size is observed. We have here kept the variables in the same order: as would be obvious to any knowledgeable party, each column references similar data within variables and different data between variables; and further, the order of variables is rarely meaningful. Also, four-digit numerical characters were used as DDIDs for simplicity alone; this referencing does not reflect the method through which Anonos BigPrivacy derives or defines its DDIDs. We assume, for conservative estimates, that we know which indicator each observation references.

Using this method, both explicit and implicit data linkages are broken, because the parsing process reassigns DDIDs to each individual observation. This effectively removes the real (explicit) *and* contextual (implicit) identifiers from the dataset, and thus eliminates the risk presented by such equivalence classes, while also masking unicity in the dataset. Specifically, it is not clear, without accessing the identification maps, whether the high blood pressure (DDID=0847) is assigned to a female person (DDID=9999). Furthermore, we cannot use that context to link data together to identify with certainty that DDID=0847 is *Alice's* blood pressure reading, as compared to DDID=6878. In this dataset with n=5 individuals, we now know that the risk of re-identification is $r = 1/n = 1/5$ (random) and since $1/5 = 0.20$, this was an acceptable level of risk. Put more strongly, any dataset with at least five observations would be compliant using the Anonos BigPrivacy method. However, even if we uniquely found a single individual with high blood pressure (as is the case above), doing so does not improve our ability to link that to any individual nor to access other information using that knowledge.

Dynamism and uncertainty

While this dataset is currently seen as time-invariant (the most conservative case for analysis, and likely when a third-party gains access to a full dataset only once, perhaps through capture of a transmission); this may not actually be the case when using Anonos' BigPrivacy method over time. Specifically, because Anonos BigPrivacy does not use temporally-static identifiers, downloading the same dataset a second time could easily lead us to reasonably, but incorrectly, conclude that the new dataset references new data because the new DDIDs are dynamic and thus new identifiers reference new data *which is also reordered*. Thus, it may be possible that the second time data are transmitted, the sample size no longer seems to reference only five people, but instead might be seen as incorporating different data and thus reference a sample that may be as large as 10. Doing so would effectively reduce the risk of re-identification so that $1/10 \leq r \leq 1/5$.

Deceptive replication

Similarly, you could mask the data in this table by adding in random information (with requisite DDIDs) and similarly mask the sample size and unicity in the data. These types of deceptive replication efforts may be further compounded because new data are randomly sorted and may incorporate more newly integrated observations. If this is the case, the duplicated or incorrect observations may *give the appearance of a larger sample size*, reducing the risk to a range ($1/10 \leq r \leq 1/5$), so that, on average assuming a uniform or normal risk distribution within that range, $r = 3/20$, a reduced risk of re-identification. However, this reduction depends on the assumption that the dataset as a whole does not contain any unique information that is not anonymized (such as the IP address above), which would be obvious once replicated and thus retain the more conservative $r = 1/n$ level of risk.

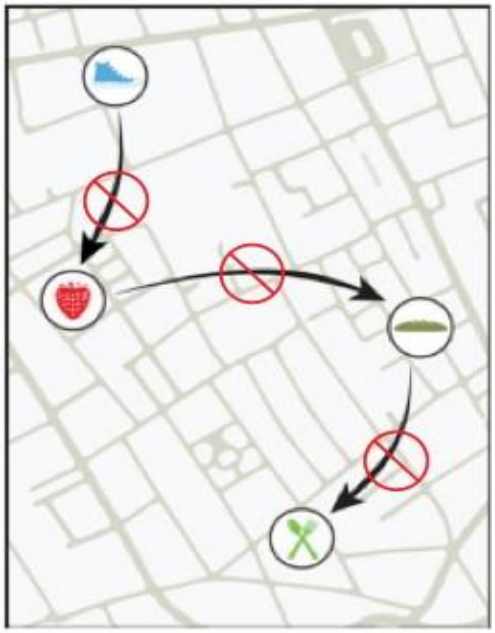
A secondary, and robust, gain from Anonos' BigPrivacy method is that we no longer know whether the two women referenced in the table above are two women or the same person measured twice. ***Challenging these assumptions can be uniquely beneficial because it forces us to question whether our basic assumptions about equivalence***

classes are correct, and further highlights the role of data that changes over time. Specifically, while we may note that here we have two people with different blood pressures, having two women the same age actually requires us to either assume that each outcome references different people (with the level of risk noted above) or to wrongly posit that these two observations reference the same person and either their health status changed over time (forcing us to question the nature of re-identification itself, since we can no longer determine when a person had the health outcome), or more likely assuming (incorrectly again) that the woman aged 42 did not have high blood pressure because there is only one observation of high blood pressure but two observations of that person.

shop	user_id	time	DDID	price
	6730G	09/23	G!022	$97.30
	S iMX	09/23	015M	$4.33
	3092fcl0	09/23	15166	$43.78
	Z3O	09/23	11A	$35.81
	4c7af72a	09/23	^99M	$15.13
	89c0829c	09/24	0099S	$12.29
	E2Grs	09/24	56KHJ	$3.66

Derived from:
Science 30 January 2015:
Vol. 347 no.6221 pp.536-539
DOI: 10.1126/science.1256297
www.sciencemag.org/content/347/6221/536

Fig. 1 Anonos Just-In-Time-Identity (JITI) enables dynamic protection at the data element level.
The universal "no symbols" highlight that dynamically obscuring data linkages that could be aggregated by parsing recognizable static "anonymous" identifiers breaks the assumptions necessary for re-identification.



Note: Figure 1 above shows that DDIDs 6730G, SiMX, Z3O and E2Grs may be used to refer to the same user at various times. And further provides DDIDs for prices, which have been randomly ordered. This obscures data linkages that could otherwise be aggregated by parsing recognizable static "anonymous" identifiers like the identifier "7abc1a23" that was used to refer to "Scott" for each transaction in de Montjoye et al. (2015).

In either case, these reasonable assumptions force us to make incorrect conclusions, making longitudinal data *less useful to re-identification than cross-sectional data*. Specifically, the risk of re-identification is no longer inversely proportional to the number of people, but instead to the number of observations: $r \leq 1/5$. This challenges the assumptions made by de Montjoye et al. (2015), who used longitudinal and specific data linkages to uniquely identify

individuals by, as is noted in Figure 1 above provided by Anonos, breaking the assumptions required for unicity. Specifically, Figure 1 notes that while multiple data-points can be used to uniquely differentiate individuals, that the Anonos BigPrivacy method breaks the implicit and explicit linkages between data points, and thus effectively removes the ability to represent data in this way. Specifically, we may know that an individual shopped at shop A, but not how much they paid at that shop nor which store they shopped at next (if any).

Put conservatively, Anonos' BigPrivacy method does not preclude the possibility that data are unique in a dataset (for example, the high blood pressure reading is unique in Tables 1 and 2 above), but makes that information useless in determining anything else about those data. It prevents unique data from being used to attempt data fusion.

Transmission

By assigning to each person a DDID, Anonos replaces individual identities with temporally dynamic random information, replacing usual static re-identification protocols that we otherwise rely on in making assumptions when re-identifying. The ability to re-identify individuals by intercepting transmitted information is predicated on data linkages – various quasi-identifiers or variables that can be aggregated by parsing the identifiers. During transmission, these data are described by Anonos as being dynamically obscured via the DDID process, and transmitted as the data available above. This corresponds to the following data stream being sent from Table 2 above: DDID=0967; High BP=No; DDID=3030; Lat.=40.29; DDID=4958; Age=42; etc.

If Anonos BigPrivacy fails, the risk that this blood pressure indicates *Alice's* blood pressure is r=1/n. However, the calculated risk is variable and, though dependent on sample size, the risk remains unknown to those interested in re-identification (because the total sample size is unknown). Moreover, deferring the transmission of this information and sending it separately increases uncertainty about the context in which the data is being delivered; because data are being de-identified, we cannot be assured, without making assumptions about this equivalency, even if only longitudinal data referencing a single person's data were being delivered over a period of time, that this data referenced multiple observations of a single person who moved around between different locations, rather than of multiple people living near each other. For example, while the above represents five people because each category provides different locations and IP addresses for a wearable dynamic, if we replaced the table above with a table referencing two women aged 45 followed up for three time points, the new data would *appear* identical to data referencing one woman aged 45 but followed up for six time points. This would be especially confusing if this person moved around between areas. Again, however, given that we know where a person was at time 1, we cannot use that information to derive information about her health or location at that time. As long as the total number of application users equals or exceeds five, and no assumptions can be made about the number and types of information available during a particular transmission (i.e., we cannot know that only one person's information is being transmitted), the risk of re-identification remains acceptable even during transmission.

Let us assume the worst-case scenario: that such a transmission was caught and *revealed in its entirety*. Then we are left with the case, explicated above, where the number of users is known. Thus, , r=1/n=1/5=0.20, which we deem to be an acceptable risk (because 0.2 ≤ 0.20). However, if the transmission is not entirely caught (for example, if blood pressure is not caught or sample sizes differ between observed variables), then the risk of re-identification must be derived from information known about the number of users of this particular blood-pressure application at this particular site. Because the number of users must be at least five (since there are five in our known dataset), we know that the risk of re-identification becomes bounded by the dataset and user-base specifics, and is thus at most 1/5 but could be as small as the actual number of potential users (n_p) so that the risk is potentially much smaller ($r=1/n_p$) since n_p is at least five but may include a huge number of users, and thus we would say that the risk of re-identification is r≤1/5.

In this case, transmission effectively replicates the "sampling" method detailed by Matthews and Harel (2011) above; a common de-identification technique in itself. Formally, this means that Anonos BigPrivacy efforts serve to increase *n* by adding to it an unknown amount (*k*), where k ϵ Z. With the addition of *k*, re-identification then relies on the probability $1/n = 1/(n' + k)$, where *k* is *unobserved*. Notably, *k* could easily be made up of data that is similar to the real data, but is fake, or by replicating randomly sorted data that is not differentiable from its copied counterpart. As a result, the risk decreases rapidly by the inverse of the total number of observed and *unobserved* users (*n*). More concretely, if we know that at *Alice's* place of work that there are 20 users of the particular application then the risk = 1/20 = 0.05, which is less than 0.20. If, however, all employees (say 350) have been provided access to the application, then the data specific risk = $1/N_{employees}$ = 1/350 = 0.0026 < 0.20. In either case, the risk is acceptable as long as the number of users in the full, accessible, dataset does not allow for $r = 1/n \geq 0.20$ (i.e., $n \geq 5$).

Optimization and efficiency

Regulatory compliance specifically requires that PD are subject to a reasonably low risk of re-identification. We have shown above that the BigPrivacy method can reduce that risk. However, it may be inefficient to dynamically identify every piece of information at all times, so understanding the necessity of such levels of security to maintaining BigPrivacy compliancy may be useful. Above, we suggested that we could parse data by cell into randomized data with unique DDIDs. However, we could maintain many of the same levels of security by viewing the main dataset as a matrix of matrices (i.e., that each matrix ***H*** contains '***j***' matrices within which the data reside). As such, DDIDs could be used to secure data not by providing DDIDs to each cell in the dataset, but instead to each matrix of cells in the dataset. This would provide much of the same level of security discussed above, but would be much more computationally efficient.

Specifically, modifying Table 1 above we provide the following dataset as a set of groups that are defined by the DDID to create Table 3 below, where each level of gray (of which there are *j*=6 made up of 3 rows and between 1 and 2 columns of data) signifies a different dataset formed of sequential or random pieces of information that could only be reassembled, like a puzzle, using the key. Here, the blocks were predefined to overlap with each type of variable in part because this overlap is easier to see and manage, but is also in many ways more secure.

Table 3. Hypothetical dataset (**H**) collected from multiple smartphones on the same network by a blood pressure application in December 2014

IP address	Latitude	Longitude	Age	Sex	High blood pressure
192.168.0.0	40.13	-79.85	32	M	No
192.168.101.201	40.13	-79.86	45	F	Yes
192.168.4.57	40.15	-79.54	39	M	No
192.168.22.40	40.29	-79.54	56	M	No
192.168.1.220	40.29	-79.56	42	F	No

It would be evident if one happened to receive one particular sub-matrix, that there was a respondent with high blood pressure (the darkest gray). However, it would be impossible to ensure that this respondent was our fictional individual, "Alice" as above. Nevertheless, it would be entirely feasible to know this if certain types of data were contained within that dataset, and thus security would only be ensured if types of data were contained separately from each other, and if the matrix mapping were not unique in itself (i.e., if matrix ***H*** could be reasonably made of by assembling these ***j*** pieces in a number of ways). Here we noted that data could be differentiated in this way: the IP address column is white because we assume it would be deleted, while the data in the blood pressure chart are held in pieces that could be made up of information in *n* (here 6) ways. As such, this provides similar security as does the method above with one caveat: if data are longitudinal and variables are stored in concert, and the outcome are sufficiently specific, then there is a small chance of matching data types together. Nevertheless, this would be reasonably solved by implementing variation in the levels of security promoted by the types of data so that publicly available data are stored in *j* cells while more sensitive data are stored in single cells without linked data. In this example, let us suggest that, under the worst-case scenario, we received the data in full but separated by shade of gray into six datasets. In this scenario, we would know because of our mapping only that one respondent had high blood pressure, resulting in a risk of re-identification of 1/5, which we have defined as acceptable. However, if this were not the worst-case scenario and only a subset of the data were received (the darkest gray box, for example) then the risk of re-identification is *at most* 1/5 and is at least $1/n_p$ where n_p includes the entire potential user base.

As noted above, this would be further secured by the occlusion tactics described above, and could be modified to secure other types of information than health data, subject to risk analysis about the specificity of that data. This application of this type of analysis has two added notes regarding specific data. Specific data could be considered to be largely "identifiable" information and would need to be separated from other forms of information to maintain privacy. Finally, longitudinal data (a form of specific data) can be stored in two ways: wide, with each variable noting an observation; or long, with each observation identified within the implicit structure underlying a dataset (as in Figure 1 above). In either case, this method could be made robust to long or wide data depending on mapping techniques and, if necessary, random sorting. Crucially, in this scenario PD would still 1) not be subject to data fusion and 2) be kept private even if unicity were achieved in the data itself.

Differential levels of security

One benefit of this type of optimization in conjunction with dynamic capabilities is that it facilitates the ability for users to manage security flexibly. Specifically, if some data were considered a greater risk than other data, it would be possible to vary the level of security used to secure different types of data and to secure data for different purveyors. For example, let us imagine that age and sex needed less security levels than a medical diagnosis of schizophrenia. It would be possible to differentiate them and use different mechanisms to organize them. Such variation matches how comfortable individuals might feel sharing information in person. For example, one could keep the age and sex variables integrated but randomly sorted. The diagnoses could, on the other hand, be differentiated and dynamically de-identified and excluded from the dataset's user base unless specifically provided by the security policy. This differentiation would both eradicate the risk of data fusion and would minimize the risk of re-identification. However, it could allow easier access to basic demographic information for accepted purposes. In this way, users could, with the guidance of policymakers *or with the explicit permission of individuals from whom they have collected data,* readily provide a sliding scale of coverage where different types of data are differentially secure.

In practice, this would imply that Anonos BigPrivacy could implement what could be termed a "programmatic policy", or a digitized representation of policy decisions that defines,* a priori*, how data are shared by specifying which data are shared when and with whom.

Unknown Third Parties

The above analyses are sufficient to ensure de-identification in the traditional, static, sense. However, we live in an increasingly demanding and dynamic world, with increasingly opaque privacy protocols. We may therefore reasonably assume that the end-user is not yet defined and that more complex associations may arise. We also may encounter the real outcome that an end-user may try to re-identify individuals in their own data surreptitiously without the knowledge of an implementation of the Anonos BigPrivacy method. We may thus be interested in knowing whether an unknown third party (U3P), not bound by data privacy standards and in possession of substantial resources (human, financial, or political), could *surreptitiously* manipulate the data to facilitate re-identification. If this is the case, then interested parties might have strong incentives to try to *find or create* a circumstance where re-identification could be made easier. These third parties may be internationally based and thus not easily dis-incentivized by standard legal considerations.

To examine this possibility, we asked the following hypothetical question: *could an end-user, with a previously specified user base, ask specific questions in order to facilitate re-identification?* Put more specifically, in an attempt to identify a target user, could a U3P modify an existing membership's data collection routine, containing the targeted user, to modify their data collection routine (but not their user base or Graphical User Interface / GUI) to clandestinely determine that user while incurring an unacceptable level of risk (> 0.20) that health data refer to a particular individual (for ease, *Alice*)?

The most readily available technique is to add questions or indicators that would easily facilitate such re-identification. For example, the risk of re-identification could be increased by defining multiple non-threatening questions that overlap in order to increase unicity and facilitate the unique identification a particular person, or by linking smartphone data or metadata (including, for example, GPS information) to publicly available information. However, because identifiable information and characteristic indicators, which might be easily added to the application in order to expressly identify the individual of interest (i.e., to maximise the risk of re-identification) are subject to Anonos' BigPrivacy method, these linkages are readily dealt with as noted above. We must therefore assume the U3P could simply access the full dataset with the identifiers from an authorized user; thus, the worst- case scenario is that they would know a person was part of the data collected. It may be possible then to gain the full

dataset, but using Anonos' BigPrivacy method, these data will not be linked and thus the result will not be more informative than that – you would know that a person was part of the data, and that there is a 1/*n* risk that any indicator, including high blood pressure, a relatively low probability. Because these data are not linked, we know that asking identifiable or characteristic questions could only be used to determine the health of a particular individual with a risk of re-identification of 1/*n*.

If identifiable or characteristic data are not useful, it may still be possible to determine/create a situation in which information is both 1) interesting in its own right, and 2) sufficiently specific to determine with risk ($r >= 0.20$) that a person fits the outcome suggested. This quest is trivial if the person of interest does not, or is not known to, use the application during the period of time under examination, since the risk will then always be $0/n = 0$. However, in our hypothetical situation, the U3P would know that the user's information was contained within the dataset provided. Thus, the data requested must, in a single unlinked variable, reference an outcome where its specificity and risk could be sufficient to identify an individual's information solely on its specifics. This is easiest when the potential outcome is strange rare (either the disease or lack of disease) since the risk of identification relies on assumptions and unicity in that dataset.

To maximise re-identification risks, we must therefore create *specific* data that are both the health variable of interest and sufficiently unique to successfully identify the information desired. This is a tall order and highly unlikely in any normal dataset, so an interloper asking these types of questions might be obvious to the respondents. In such a situation, we might reasonably assume most individuals to be free of the disease, and that we have reason to believe that the risk that *Alice* has the disease is *M* times greater than the normal population. Nevertheless, we want then to know what the likelihood is that *Alice* (*A*) has condition *R*, given that *R* is observed in the dataset (denoted, $P(A|R)$). This calculation can be solved using Bayes' Theorem. The probability *Alice* has the disease is: $P(A|R) = P(R|A)*P(A)/P(R)$. These other probabilities are either known or can be guessed. For example, the probability that any observation is *Alice's* is $P(A) = 1/n$. The probability that any sample contains an observation of that is P(*R* particular disease) $= (R*(n-1)+M*R)/n = R*(n-1+M)/n$, where *R* (such that $0 \leq R \leq 1$) is the risk of the disease. We believe, from our expectations derived from external observation, that if *Alice* has a risk *M* times the normal risk (*R*) of observing the outcome such that $Q = 1-R$, then the probability of a positive outcome given that *Alice* is in the sample is $P(R|A) = MR$. Thus, we have from Bayes' Theorem that $P(A|R) = P(R|A)*P(A)/P(R) = (MR*1/n)/(R*(n-1+M)/n) = M/(n-1+M)$.

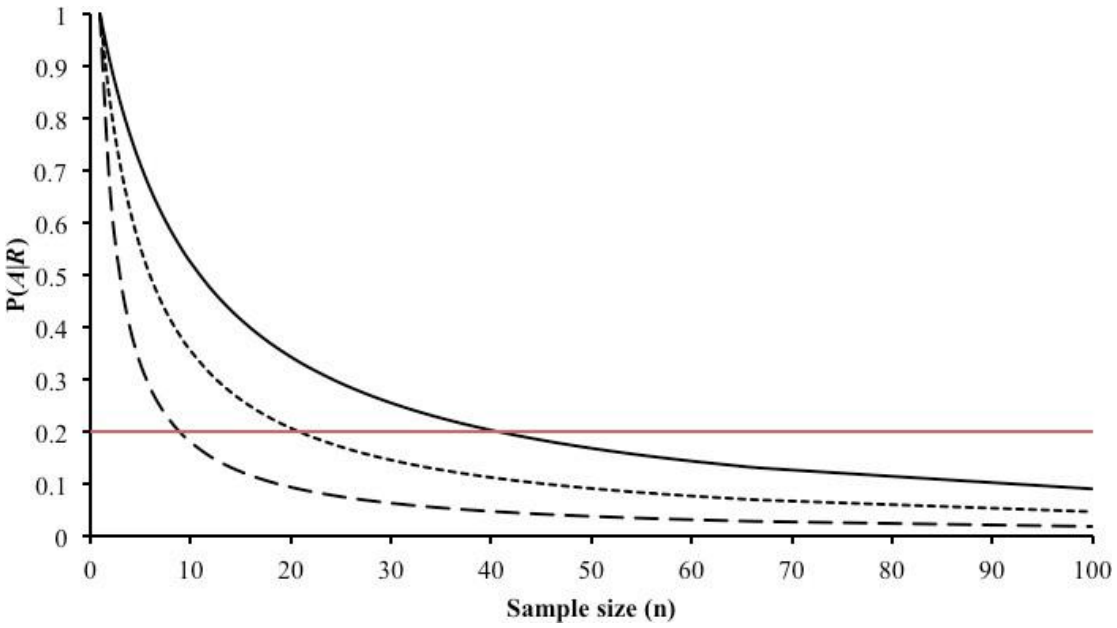


Figure 2. Likelihood of re-identification under the hypothetical condition that data are manipulated in order to engineer the best conditions possible to identify individuals, with an estimated *M* of 2 (long dashes), 5 (dotted), and 10 (solid) by sample size.

Simulations estimating the risk of re-identification given a single positive observation follows Figure 2 above. We have here assumed a range of relative risks ranging from conservative (*M* = 2) to medium (*M* = 5) to very large (*M* = 10). This range of relative risks (*M*) was allowed to range from 2-to-10 to reflect the range often seen for predictors in epidemiological research, and because at most *M* references the difference between a risk of 2% (a rare outcome) and 20% (the risk necessary to be reasonably certain = 0.20). Notably, risks much higher than 2% become decreasingly likely to enable an *M* = 10 outcome because when the population's risk approaches 10%, the personal risk must approach 100% (i.e., 10*10%), a known certainty, and thus the need for re-identification is unnecessary to begin with.

Figure 2 above provides a more conservative estimate of the risk of re-identification than traditional methods. This estimate suggests that in the worst possible situations that Anonos' BigPrivacy method is robust to intentional privacy intrusions by a U3P undertaken with express knowledge of the Anonos BigPrivacy method, as long as total sample size exceeds 41 individuals (the point where the solid black line (*M* = 10) crosses = 0.20). Notably, while it is unlikely that all data are going to need this level of privacy, it is reasonable to suggest that when data are treated in this manner that they achieve or surpass this stringent level of security.

Discussion

In this analysis, we described Anonos' BigPrivacy method as starting with the premise of blending existing methods of de-identification, including for example sampling, suppression and the potential addition of noise, with novel temporally dynamic identifiers and data parsing protocols. We have analysed the risk of re-identification, finding that the Anonos BigPrivacy method can drastically reduce the risk of re-identification, even for specific data. Moreover, these analyses we found that, using the Anonos BigPrivacy method, data were kept private during both transmission and storage, even from the application developers. Specifically, we found that re-identification risks were minimized and could be reduced under the generally accepted statistical and scientific principles and methods for rendering information not individually identifiable (threshold value (here defined as = 0.20)), when the total sample size equalled five analytic units (e.g., individuals, households, online identities, IP addresses, etc.). Moreover, we discussed the potential for BigPrivacy processes to be applied to blocks of data rather than individual observations. We further found that the level of security could be managed by variable differentiation and de-linkage, so that some information, such as basic demographic information was not de-identified but other information was at the same time subjected to the BigPrivacy process. We further discussed the potential for both policymakers and for individuals from whom the data are collected to help define the level of security of particular data.

Risk of re-identification

The risk of re-identification is limited by constraining assumptions that can be made about data contents and structure (Kifer & Machanavajjhala, 2011). *Anonos works by breaking the assumptions that are encoded in datasets and used by others to achieve re-identification.*

Breaking these assumptions has a number of benefits, but the most important one is that it makes the both re-identification and the risk of re-identification difficult to ascertain with any level of certainty without further gaining access to the complete, unadulterated, dataset. Anonos BigPrivacy does this in a few main ways discussed below.

Being dynamic helps. DDIDs provide a level of protection from data and the misuse of data that are not available now. For example, DDIDs necessarily re-integrate randomized follow-up information from the same individuals if data were downloaded later, and thus serve to increase sample size and reduce re-identification risks while reducing our ability to make assumptions about the completeness of the data. Secondly, the data differentiate instances from one another, making assumptions about the completeness of data, and their reference population, less clear. Third, Anonos BigPrivacy efforts work well during transmission to effectively occlude shared information and to maintain security even with characteristic and specific data. Finally, the method can be made robust even to those who are engaged in collecting the data, making data privacy clear and enforcing data use agreements even when unknown third parties are engaged in using the data.

ANONOS®

Flexibility of privacy and security

This technology enforced decoding of DDIDs could apply broadly, within a single deployment of the Anonos BigPrivacy method, but it would be possible to overlap multiple cascading rule sets, with an agreed-upon hierarchical relationship, to govern usage of any given primary data table. In practice, this could mean that a lead country's Data Protection Authority (DPA) might define the highest-ranking set of PD access rules, but another concerned party might also insert its own set of PD access rules that may be more stringent. These rules might be applied differently to different types of data within the same dataset. In this event, Anonos can be configured to ensure that no PD access is possible unless both cascaded sets of DPA access rules are enforced when the query is made. Conversely, BigPrivacy could provide flexible controls necessary to support hierarchical handling of various data privacy requirements.

Programmatic policy

The ability to deliver on the many promises of big data in linking together individuals with institutions, clinicians, or researchers, for example, is predicated on this ability to support differing privacy requirements depending on the nature and source of data. Anonos BigPrivacy provides a way to automatically and digitally enforce such privacy policies. For example, consumer health data collected using electronic health records, mobile health applications, and social networking sites may be accessed and data may be useful and available. At the same time, financial data may be transcribed into the same data using the same devices. However, PD is at the same time regulated by privacy and security requirements under a given country's privacy and health privacy acts and may further be subject to specific privacy policies and terms and conditions depending on user preferences for specific websites, devices and applications. The BigPrivacy key itself encodes both the rules necessary to recover the source value from at least one DDID and flexible programmatic policies, or a digitized representation of the privacy policy that is subject to observation, enforcement, and audit. Therefore, if necessary rules and constraints are not being observed and enforced, either because there is 1) a mismatch between a user's permissions and the query that user is trying to submit or 2) access was once granted but has since been revoked or expired, then no DDIDs may be decoded.

Conclusion

The Anonos BigPrivacy invention and protocol mitigates the risk of re-identification by repudiating assumptions about explicit and implicit data linkages. It can therefore ensure privacy even when the dataset as a whole contains characteristic or specific data, such as when single individuals are followed over time or specific details such as geographic location are observed.

The flexibility of the BigPrivacy key mechanism ensures that the Anonos policy-driven BigPrivacy data management platform, even in cases where a single data element of PD must be protected by a single BigPrivacy key, programmatically enforces granular rules for access. We also found that, even when individuals worked to design a situation favouring re-identification, Anonos' BigPrivacy method continued to minimize the risk of re-identification by first removing the risk of characteristic re-identification, while repudiating the ability to make assumptions about the structure of the data, and also by limiting the risk of specific re-identification to acceptable levels given sample size limitations. We then identified opportunities to both: further occlude data in cases of small numbers of observations, and optimize occlusion to facilitate large-scale data management.

It is the author's opinion from the analyses conducted and described herein that, subject to oversight and sample size limitations, Anonos' BigPrivacy method substantially mitigates to a statistically acceptable level the ability to single out, infer about, or link data to an individual so that personal data remains private.

Author Biosketch

Sean Clouston, PhD, is a *Fulbright* scholar who earned his Bachelor's in Arts in Mathematics and Sociology. He did his doctoral work at *McGill University* in statistics, epidemiology, and demography and is currently employed as Core Faculty in the Program in Public Health, and an Assistant Professor of Family, Population, and Preventive Medicine at *Stony Brook University*, part of the State University of New York. Dr. Clouston uses quantitative analysis on high dimensional data to examine questions relating to the distribution and determinants of disease both in the United States and globally.

ANONOS®

References

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association, 17*(2), 169-177.

Berinato, S. (2015, February 9th). There's no such thing as anonymous data. *Harvard Business Review*.

Castro, D. (2015). FTC's Internet of Things Report Misses the Mark [Press release]

Chen, D., & Zhao, H. (2012). *Data security and privacy protection issues in cloud computing.* Paper presented at the Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on.

Cox, L. (1994). Matrix masking methods for disclosure limitation in microdata. *Survey methodology, 20*(2), 165-169.

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science, 347*(6221), 536-539. doi:10.1126/science.1256297

El Emam, K. (2015). Is it safe to anonymize data?

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., & Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian journal of hospital pharmacy, 62*(4), 307.

Federal Trade Commission. (2015). *Internet of things: Privacy & security in a connected world*. Washington, DC: Federal Trade Commission.

Herz, J. (2014). Wearables are totally failing the people who need them most. *Wired*.

Hotz, R. L. (2015, January 29). Metadata Can Expose Person's Identity Even When Name Isn't; Researchers Use New Analytic Formula. *Wall Street Journal*.

Kifer, D., & Machanavajjhala, A. (2011). *No free lunch in data privacy.* Paper presented at the Proceedings of the 2011 ACM SIGMOD International Conference on Management of data.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys, 5*, 1-29.

Measure DHS, & ICF International. (2013). *Demographic and Health Surveys*. MD5.

Rivera, J., & van der Meulen, R. (2014). Gartner says the Internet of Things will transform the data center. *Gartner*. Retrieved from http://www.gartner.com/newsroom/id/2684616

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 461-468.

Tyrrell, C. (2014). Countering HITECH privacy risks from internet of things products. *HITECH*

Answers. Warren, N. (2014). Dynamic Data Obscurity. *Category Archives.*